uc3m | Universidad Carlos III de Madrid

Grado en Ingeniería Informática

2016-2017

Trabajo Fin de Grado

Clasificación de ondas sísmicas con técnicas de minería de datos

---

María de las Mercedes Crespo Jiménez

Tutor

Agapito Ledezma Espino

Leganés, Octubre 2016

Agradecimientos

Quiero agradecer a mis padres y a mi hermana todo el cariño, consejos y apoyo que me han dado en todo momento a lo largo de mi vida porque gracias a vosotros la felicidad es una realidad a día de hoy, con toda mi vida por delante, puesta a mi alcance. Me habéis enseñado a ser agradecida y no desfallecer y habéis sido mi hogar, mi lugar seguro en el mundo.

Sois excepcionales.

Gracias.

Agradecer a todos los profesores que me han ayudado a acabar el Grado, con especial cariño a los profesores del campus de Colmenarejo: Miguel Ángel Patricio, Antonio Berlanga y Jesús García. También a los profesores Gonzalo Génova, José Daniel García, Dolores Cuadra y a Alfonso Martos Abascal.

Este trabajo no podría haberse realizado sin vuestra colaboración profesional y personal a lo largo de mi vida académica.

Mención aparte a mi tutor Agapito Ledezma Espino puesto que fue el único que apostó por mi idea y me facilitó todas las herramientas a su alcance para que la llevara a cabo y por supuesto a José Manuel Molina López porque sin tu apoyo, críticas y reflexiones ni la elaboración de este trabajo ni mi experiencia académica habría sido la misma.

Gracias.

Por último, agradecer a todos mis amigos su cariño desinteresado hacia mí, en especial a mis compañeros de Leganés y a Marcell Albuín Álvarez.

Gracias.

Resumen

Este proyecto surge de mi interés personal por los terremotos, causas y consecuencias.

Tengo una naturaleza sensible y la destrucción de Fukushima en el año 2011 me produjo una fuerte impresión que ha permanecido inalterable hasta el día de hoy.

El orden, la solidaridad y la entereza del pueblo japonés frente a la catástrofe me conmovieron. A diferencia, de lo que hubiera cabido esperar, los japoneses no realizaron robos, asesinatos o violaciones; se agruparon, ayudaron y reconstruyeron sus vidas, algunos de ellos prácticamente solos.

La prevención de terremotos es una tarea pendiente sobre la que se ha realizado una gran investigación con apenas avances hasta hoy. La capacidad de la que se dispone en la actualidad para almacenar datos y explotarlos no tiene ni punto de comparación con la del siglo pasado. Las técnicas de Minería de Datos serán decisivas en los próximos años para realizar análisis de la actividad sísmica de distintas partes del mundo y realizar nuevos experimentos y plantear hipótesis que puedan ayudar a prevenirlos de manera efectiva.

Este trabajo tiene como objetivo proponer una alternativa a un proceso manual dentro de esta área: la clasificación de ondas sísmicas a partir de los registros facilitados por un conjunto de estaciones sismográficas.

A partir de los datos de la monitorización volcánica del volcán Puracé en el período que abarca Julio de 2015 a Julio de 2016 facilitados por el Instituto de Geología y Minas de Colombia se estudiará la posibilidad de realizar clasificadores de ondas sísmicas teniendo como única entrada esta información.

Tras realizar la correspondiente investigación sobre el dominio, parte de estos datos serán seleccionados y transformados para poder ser empleados por la herramienta de Aprendizaje Automático y Minería de Datos, WEKA. Con ellos, se elaborarán modelos de clasificación y sus resultados se acompañaran de una serie de experimentos de segmentación que refuten la validez de los resultados obtenidos y se obtenga información desconocida del dominio.

Por último, se añadirá un apartado con las conclusiones de este trabajo en el que se valorará si el objetivo propuesto se alcanzó y se plantearán futuras líneas de trabajo relacionadas con este proyecto y la Inteligencia Artificial.

Abstract

Due to the impact that Fukushima's tsunami in 2011 had on me I wanted to focus this Project on seismic activity.

Japanese people behave after the nuclear disaster with strength and solidarity, helping each other and challenging the uncertainty of their future with bravery and hope. Their attitude brought to tears and desperations, how many rotten lives must have continued after that? Did we do everything we could to prevent the disaster?

Even if we did, don't we must try harder?

These questions brought me here. Earthquakes forecasting is a challenging idea after which we have been running after for several centuries. Machine Learning and Data Mining techniques can help to this purpose.

The aim of this project is creating accurate seismic wave's classifiers with Data Mining and Machine Learning techniques that may prove their effectiveness in this domain.

Using the volcano's monitoring data of Puracé's volcano seismographs from July 2015 till July 2016 several classifiers would be created using this information as the only one entry of data.

Some of this data will be selected and prepared to be used by the Machine Learning tool WEKA. WEKA will output classifiers and also the results of some clustering experiments that may give unknown information about the domain and may help backing up classifier's results.

Some conclusions will be added at the end of the Project evaluating the performance of all the classifiers and suggesting future work lines.

Tabla de contenido

Índice de Figuras

Índice de Tablas

1. Introducción

Desde que el hombre es hombre ha tratado de explicar el mundo. Empleando los sentidos o el intelecto trata de controlar su entorno y en última instancia, explicarlo.

Las primeras sociedades prehistóricas surgieron con la aparición del lenguaje y su capacidad para crear y razonar sobre elementos abstractos.

Despojados del confort de nuestros días, la humanidad no buscaba entonces un porqué o un sentido a su existencia, se conformaban con perpetuarla tanto como fuera posible. Frente a su incapacidad para explicar los fenómenos que ocurrían a su alrededor, surgieron los primeros ritos y creencias.

Siglos después, la ciencia ha permitido explicar algunos de estos mitos que en su origen eran atribuidos a dioses o en el caso de los cultos animistas a las almas o espíritus de los objetos de su entorno y paralelamente a su desarrollo, se iban creando y destruyendo distintas sociedades.

La humanidad en su afán de control y búsqueda de seguridad ante un mundo que no entiende ha ido desarrollando complejas estructuras sociales, fuertemente antropocéntricas, entre las que destacan la economía y la política con la que se continúa sometiendo a unos y encumbrando a todos y explotando de manera indiscriminada los recursos naturales.

Y sin embargo, este poder es un espejismo ante fenómenos como incendios, tornados o terremotos. Estos últimos se han multiplicado en la última década. Impredecibles y destructivos arrasan vidas, pueblos, países. Desde la Antigua Grecia hasta nuestros días se han ido registrando, cuantificando y midiendo sus daños, midiendo su frecuencia y ampliando el conocimiento que se tiene sobre ellos.

Sismólogos de todo el mundo analizan la actividad sísmica de volcanes y puntos de alta actividad sísmica y tratan de averiguar cuándo será el siguiente, qué valor tendrá su magnitud, cuántas vidas podrían salvar si consiguieran predecirlo a tiempo.

Parte de este análisis es la clasificación de las ondas sísmicas liberadas durante un terremoto. Estas ondas a día de hoy se clasifican de manera manual privando de un tiempo precioso de investigación a los especialistas. La minería de datos permitiría realizar esta tarea mediante la creación de modelos de comportamiento hasta ahora desconocidos.

La actividad sísmica del volcán colombiano Puracé, comprendida entre julio de 2015 y julio de 2016 será empleada para elaborar modelos de clasificación de ondas que permitan identificar con la mayor fiabilidad posible entre los 18 posibles tipos de sismos de los que a día de hoy se tiene constancia.

Este documento inicia con una breve introducción sobre la minería de datos y el análisis de datos en la que se abordarán sus dominios de aplicación y relación con otras disciplinas para a continuación, ahondar en su relación con la sismología incluyendo finalmente la experimentación y resultados de las técnicas empleadas

sobre los datos de estudio que comprenden experimentos con técnicas de clasificación y de segmentación.

El marco legal en el que se enmarca el proyecto se recoge en el punto 1.4 y el entorno socio-económico en el apartado 7.

1.1 Motivación

Dada la potencial amenaza que entraña la actividad sísmica para el planeta, la idea de realizar un Trabajo de Fin de Grado que abordara esta temática y permitiera estudiarla desde el punto de vista de la Inteligencia Artificial me resultaba atrayente puesto que apenas hay publicaciones y estudios que relacionen ambos temas.

1.2 Objetivo

El propósito general de este Trabajo de Fin de Grado es el análisis de la actividad sísmica del volcán Puracé elaborando clasificadores de ondas sísmicas mediante técnicas de Aprendizaje Automático.

1.3 Entorno legal

Los potenciales problemas legislativos se enmarcan en el ámbito de la gestión de información. Estos datos fueron extraídos de un repositorio [1] de libre acceso perteneciente a la Universidad Carlos III de Madrid pero al desconocer si al emplear los datos se puede incurrir en algún tipo de delito ni falta, se realizó una consulta a la página web de la Agencia Española de Protección de Datos desde la que se tiene acceso a la legislación vigente en todo el territorio nacional.

En primer lugar se consulta la Ley de Protección de Datos (Datos 1999) [2] que actualmente se aplica en todo el territorio español. En el Artículo Tercero del Título Primero de esta ley se encuentra la definición de *datos de carácter personal* que dice:

"Datos de carácter personal: cualquier información concerniente a personas físicas identificadas o identificables."

Estos datos pertenecen al Observatorio Geológico de Colombia que posteriormente los compartió con el Grupo de Control, Aprendizaje y Optimización de la Universidad Carlos III, el cual los pre-procesó y publicó en su página web con fines de investigación. Si bien es posible identificar la propiedad de estos datos, en ningún caso es posible identificar a las personas físicas que los obtuvieron y manipularon, por lo que será posible utilizarlos sin incurrir en falta o delito siempre y cuando estén referenciados a la entidad a la que pertenecen.

[1] https://www.caos.inf.uc3m.es/datasets/

[2] http://www.agpd.es

En el Título Segundo Artículo Cuarto Punto Segundo de esta ley se indica claramente que el análisis estadístico de los datos sí está protegido por la propia ley que dice:

"Los datos de carácter personal objeto de tratamiento no podrán usarse para finalidades incompatibles con aquellas para las que los datos hubieran sido recogidos. No se considerará incompatible el tratamiento posterior de éstos con fines históricos, estadísticos o científicos."

El propósito de este proyecto hace imposible atentar contra ella y sus resultados también estarían protegidos siempre que no se empleen con fines comerciales al poder ser considerados como un estudio con fines científicos.

La lectura completa de la ley lleva al Artículo Seis Punto Segundo que versa sobre la autorización de explotación de los datos y dice:

No será preciso el consentimiento cuando los datos de carácter personal [...] cuando los datos figuren en fuentes accesibles al público y su tratamiento sea necesario para la satisfacción del interés legítimo perseguido por el responsable del fichero o por el del tercero a quien se comuniquen los datos, siempre que no se vulneren los derechos y libertades fundamentales del interesado.

Tal y como se especifica en la página web, el conjunto de datos que se utiliza se va a emplear para comprobar si es posible emplear algoritmos de aprendizaje automático que puedan generar modelos de clasificación de ondas a partir de los datos disponibles.

Basándose en los artículos arriba citados se concluye que se ha respetado la legalidad vigente en el país en todo momento.

2. Estado de la cuestión

2.1 Introducción

La sismología es una de las disciplinas que abarca la geología, el estudio de la Tierra, concretamente la rama de la geofísica que se encarga del estudio de ondas mecánicas que se generan en la corteza terrestre (Minería s.f.).

Los documentos más antiguos de esta disciplina data de la Antigua Grecia produciéndose escasos avances hasta el siglo XVII, momento en el que se produjeron grandes avances en la disciplina de la física que permitieron a los científicos de la época desarrollar experimentos que plantearan hipótesis hasta ahora impensables ante la falta de conocimiento.

Un siglo después, en 1775, tuvo lugar uno de los terremotos más devastadores jamás registrados, el terremoto de Lisboa, que coincidió con el florecimiento general de la ciencia en Europa y disparó el interés científico por comprender el comportamiento y la causa de los terremotos. Los ingleses John Bevis (1757) y John Michell (1761) realizaron trabajos de gran importancia en el campo de la sismología. En 1757, Bevis publicó en Londres *The History and Philosophy of Earthquakes* libro en el que se documentaba el terremoto de Lisboa a partir de distintas fuentes. Su trabajo fue posteriormente utilizado por John Michell quién determinó en 1761 que los terremotos son ondas de movimiento causadas por "masas de roca que se mueven millas por debajo de la superficie" de la Tierra. Por otro lado, el gobernador de Portugal, el Marqués de Pombal inició la reconstrucción de la ciudad y llevó a cabo estudios científicos que documentaron las repercusiones del terremoto y en el que participaron científicos de otras partes del mundo. Fue el primer estudio llevado a cabo con colaboración internacional (Chatfield 2010).

En 1857, Robert Mallet fundó la sismología instrumental y llevó a cabo experimentos sismológicos utilizando explosivos. En 1906, Richard Dixon Oldham identificó el arribo separado de las ondas P, las ondas S y las ondas de superficie en los sismogramas, y además encontró una evidencia clara de que la Tierra tiene un núcleo central de una composición que le es propia.

En 1906 tuvo lugar otro terremoto devastador, el terremoto de San Francisco que, nuevamente, atrajo el interés de la comunidad científica. De nuevo, los avances en otras disciplinas fueron de gran importancia a la hora de que Harry Fielding Reid pudiera elabora la teoría del rebote elástico, la cual sigue siendo la base de los estudios tectónicos modernos.

Esta teoría establece que las rocas en el interior de la tierra soportan grandes tensiones hasta ver superada su constante de deformación elástica momento en el que liberan toda la energía acumulada y se resquebrajan al tratar de recuperar su forma original.

Alfred Wegener a principios del siglo XX propuso la teoría de la deriva continental para explicar el hecho de que los continentes pareciesen piezas que encajaran entre sí, como si se trataran de un puzle. Sugirió la existencia en el pasado de un único continente llamado Pangea, que se habría fragmentado en el período Jurásico y habría

dado lugar a los continentes actuales. Sin embargo, Wegener no fue capaz de argumentar cómo había tenido lugar esta separación y su teoría no fue retomada hasta después de su muerte.

John Tuzo Wilson (1908 – 1993) propuso la teoría de expansión del fondo oceánico basándose en observaciones geológicas y geofísicas que indicaban que las cordilleras meso-oceánicas funcionan como centros donde se genera nuevo piso oceánico conforme los continentes se alejan entre sí, es decir, que se renovaba y esos cambios podían tener repercusiones en la superficie, en la forma de volcanes y terremotos.

En base a las teorías de Wegener y Tuzo Wilson nació la teoría de tectónica de placas producto de la colaboración internacional en el que geólogos (Tuzo Wilson, Walter Pitman), geofísicos (Harry Hammond Hess, Allan V. Cox) y sismólogos (Linn Sykes, Hiroo Kanamori, Maurice Ewing) colaboraron.

2.2 Sismología

Los objetivos de esta ciencia son:

1. Conocer la estructura interna de la tierra.
2. Estudiar las causas que originan los movimientos sísmicos.
3. Prevenir daños.
4. El estudio de otros fenómenos asociados como maremotos, tsunamis y dinámicas volcánicas.

2.2.1 Definición de terremoto

Un terremoto es un evento sísmico originado en las proximidades de una falla a cierta profundidad bajo tierra en un punto determinado llamado foco o hipocentro. Este punto recibe el nombre de epicentro en la superficie.

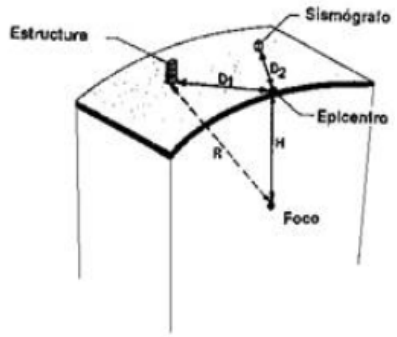


FIGURA 1 - Puntos singulares de un terremoto[3]

Este evento sísmico tiene lugar cuando los dos extremos de una falla entran en contacto produciéndose un ligero encajamiento entre ambos en el que las rocas de un lado oprimen las rocas del otro lado. Esta presión permanece en el tiempo y las rocas oprimidas sufren una deformación elástica de la que se liberan fracturándose y

[3] http://www.educiencias.es/sismologia.html

liberando la energía acumulada de manera repentina, provocando las ondas sísmicas que se perciben en la superficie y causan los terremotos.

La actividad sísmica también se puede deber al factor humano. Las explosiones bajo tierra o la construcción de carreteras pueden provocar ondas mecánicas que pueden generar ondas sísmicas débiles, a menudo imperceptibles en la superficie. Sin embargo, también se pueden provocar ondas sísmicas fuertes por la denotación de una bomba nuclear.

2.2.2 Herramientas actuales

Los sismólogos estudian los terremotos empleando sismógrafos y observando *a posteriori* sus efectos. Un sismógrafo es un instrumento que recoge los temblores producidos por ondas sísmicas. Se localizan bajo tierra y a menudo se disponen en redes. Los sismógrafos se colocan en zonas de actividad sísmica y esto implica que en ocasiones la instalación se lleva a cabo en los propios hogares como es el caso de California (Seismic Systems 2008). En el lenguaje natural las palabras sismógrafo y sismómetro se utilizan indistintamente pero en realidad, el sismómetro es tan solo la parte interna del sismógrafo.

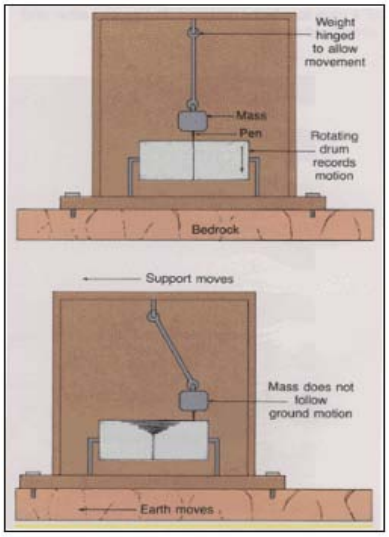


FIGURA 2 - Sismógrafo (LUTGENS & TARBUCK, 1989)[4]

[4] http://www.geo.mtu.edu/UPSeis/studying.html

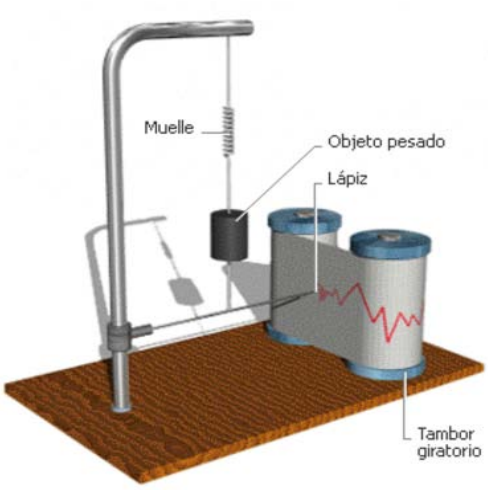


FIGURA 3 - Descripción de un sismógrafo empleando una herramienta de diseño gráfico[5]

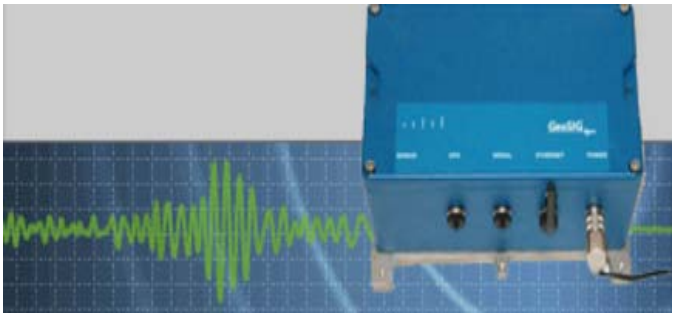


FIGURA 4 - Sismógrafo Moderno[6]

2.2.3 Escalas

Las escalas empleadas para medir los efectos de un terremoto son las escalas de intensidad y las escalas de magnitud.

Las escalas de intensidad miden los desperfectos causados por un terremoto basándose en los testimonios de víctimas y testigos, por lo que sus mediciones no se consideran del todo fiables al estar impregnados inevitablemente de subjetividad. En función del lugar del mundo en el que se registre el terremoto, se emplea una u otra distinta escala de intensidad por ejemplo, en China se emplea la escala de Liedu, en Europa, la escala macro-sísmica europea, en Hong Kong y Estados Unidos la escala de Mercalli Modificada, en India, Israel, Kazajstán y Rusia la escala Medvedev–Sponheuer–Karnik, en Japón y Taiwán la escala Shindo y en Filipinas la escala de intensidad PHIVOLCS (PEIS).

[5] http://geofisicasismospgf.blogspot.com.es/p/sismografo.html

[6] http://www.seismicsystems.net/

Todas ellas se basan en la escala de Mercalli inventada por Giuseppe Mercalli en 1902.

FIGURA 5 - Giuseppe Mercalli[7]

Las escalas de magnitud en cambio emplean las medidas tomados por instrumentos fiables como por ejemplo, sismógrafos, y tienen una base científica.

Existen varias escalas siendo las más conocidas la escala de Richter y la escala sismológica de magnitud de momento (M_w).

La escala de Richter fue inventada por Charles F. Richter en 1934 para medir la actividad sísmica del sur de California a partir de los datos de las estaciones sismográficas cercanas. Los datos que se obtenían correspondían a ondas sísmicas de alta frecuencia y eran registradas por un tipo de sismógrafo concreto, que se empleaban posteriormente para calcular la magnitud de Richter. Esta magnitud era interpretada por Richter quién le daba un valor que medía la intensidad del terremoto. Posteriormente, realizó una tabla que permitía su clasificación directa.

FIGURA 6 - Charles Richter y un sismograma[8]

[7] http://www.geo.mtu.edu/UPSeis/intensity.html

[8] http://www.geo.mtu.edu/UPSeis/intensity.html

La escala de Richter empezó a ser adoptada a nivel mundial pero se observó que sólo realizaba análisis precisos de determinadas frecuencias y un único tipo de onda.

En base a la metodología con la que elaboró su escala, se desarrollaron otras basadas en la medición de otras magnitudes como las escalas M_d, M_c o M_{ms}, siendo posible en todos los casos, establecer una equivalencia con la escala de Richter y diferenciándose de ella en el tipo de señal sísmica que podía registrar y la frecuencia estudiada. Sin embargo, ninguna de estas escalas se podía emplear a nivel global, pues sólo trataban una única onda, un único tipo de actividad sísmica. Aún no existía una escala que se pudiera emplear a nivel global y cubriera varios tipos de onda.

En 1979, Thomas C. Hanks y Hiroo Kanamori introdujeron la escala sismológica de magnitud de momento, en su forma abreviada M_w, capaz de medir la intensidad de un terremoto basándose en la magnitud del momento. El momento sísmico es el producto entre el área donde ocurrió el terremoto, su desplazamiento aproximado y el módulo de deformación (una constante elástica) de las rocas. Este resultado se convierte a un numero interpretable por una escala empleando una formula y el resultado es la magnitud del momento sísmico. Esta escala puede medir cualquier tipo de onda en cualquier parte del mundo.

2.2.4 Lectura de sismogramas

En un sismograma se recogen las ondas sísmicas provocadas por un terremoto. La lectura de un sismograma se realiza de izquierda a derecha. Su eje horizontal determina el tiempo (medido en segundos) y su eje vertical, el desplazamiento del suelo (medido en milímetros). La mayoría de los sismogramas actuales son digitales, pero aún quedan sismógrafos analógicos que registran la información sobre papel o cintas.

Un sismograma o registro de actividad sísmica se compone de cuatro elementos: pre-evento, onda P, onda S, ondas de superficie, coda.

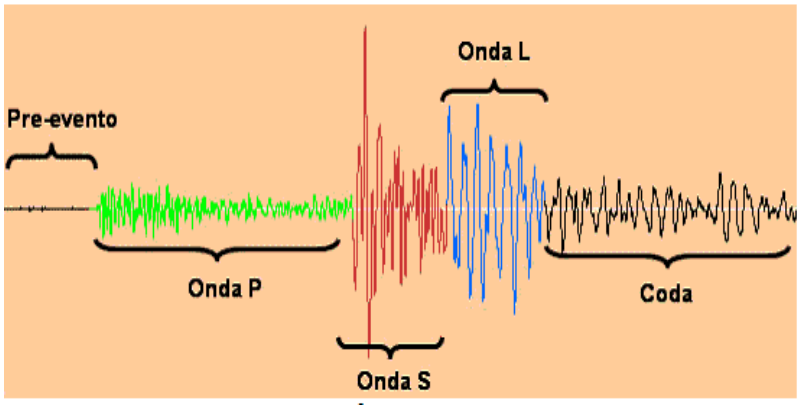


FIGURA 7 - Partes de un sismograma[9]

[9] http://www.lis.ucr.ac.cr/index.php?id=7

Cuando un terremoto ocurre en el interior de la Tierra se libera energía que viaja a través de la Tierra en forma de ondas sísmicas con distinta fuerza y velocidad.

Cuando un sismo va a ocurrir, se recogen pequeños temblores prácticamente imperceptibles que reciben el nombre de pre-evento que constituye una pequeña parte del sismograma. Inmediatamente después, llega la primera onda, la onda P u onda comprensiva de longitud variable. Es la onda más rápida y por lo tanto, la primera en llegar a un sismógrafo. Esta onda comprime y expande las rocas en la misma dirección en la que se propaga. Es la más rápida y viaja en dirección paralela a la dirección de propagación de la onda sísmica atravesando medios sólidos y líquidos.

La onda S u onda de cizalla, es la onda inmediatamente posterior a la onda P y en este caso se propagan en dirección perpendicular a la dirección de propagación. Solo atraviesan medios sólidos. En caso de no existir ondas S en el sismograma, es probable que el terremoto haya ocurrido en otro punto de la Tierra, pues, estas ondas no pueden propagarse por las capas líquidas de la Tierra y por tanto, no pueden ser detectadas por el sismómetro porque se han extinguido.

Las ondas P y S son las ondas que se propagan en el interior de la Tierra y se las denomina ondas de cuerpo. Las ondas posteriores, son las ondas que se propagan por la superficie y reciben el nombre de ondas de superficie o superficiales. Los dos tipos más comunes son las ondas Love y Rayleigh. Estas ondas ocupan la parte central del sismograma.

Las ondas L y R son más grandes y lentas que las ondas S (a su vez más lentas y grandes que las P) y su frecuencia es inferior a las de P y S lo que se traduce en un una mayor amplitud en el trazo, mayor cuanto más próximo esté el foco a la superficie. Estas ondas mueven la superficie en distintas direcciones. Las ondas Love la mueven de lado a lado en un plano horizontal, a ángulos rectos de la dirección de propagación y las ondas Rayleigh en los planos vertical y horizontal siguiendo una forma elíptica.

La última parte del sismograma es la coda que recoge el decaimiento del temblor hasta volver a ajustarse a los valores del pre-evento. La duración de la coda varía en función del tamaño del sismo.

La duración de un sismo consiste en medir el tiempo que dura la señal, tomando como el inicio la llegada de la primera onda P hasta que termina la señal. Además de la duración del sismo, otros elementos pueden ser detectados en un sismograma como la amplitud y el período. La amplitud es el máximo desplazamiento vertical de la onda sísmica (de arriba a abajo). Se puede medir de dos maneras: simple o pico a pico. La medida de la amplitud pico a pico se calcula en milímetros midiendo desde el pico máximo hacia abajo hasta el pico máximo hacia arriba en la parte donde la onda tenga su mayor amplitud o pico. La medida simple de la amplitud consiste en la medida tomada a partir de la línea base o centro de la señal sísmica (en la vertical), el máximo pico de la onda hacia abajo o hacia arriba. El período consiste en medir la duración de un ciclo completo de una onda sísmica a partir de la línea base. En la imagen que se adjunta (véase Figura 8) se observa cómo localizar y medir estos elementos dado un sismograma:

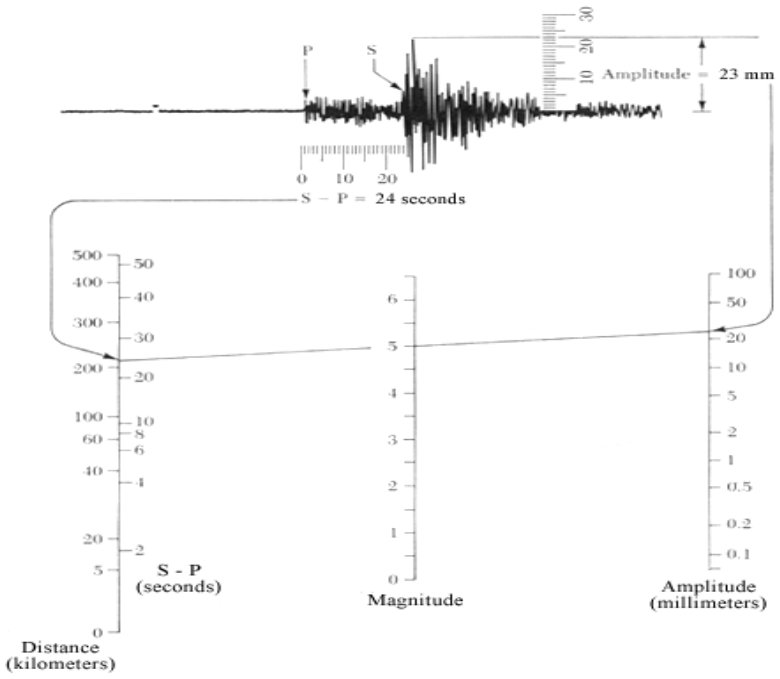


FIGURA 8 - Sismograma y elementos característicos (Bolt, 1978)[10]

2.2.5 Localización del epicentro de un terremoto

La lectura de un sismograma permite establecer la hora en que tuvo lugar cierta actividad sísmica y el epicentro (origen en la superficie) interpolando los resultados de varios sismógrafos que hayan registrado la misma actividad sísmica. Según la ubicación del sismógrafo, la hora señalada será distinta y será necesario interpolarla para dar con la hora exacta. Al margen de este detalle, el proceso de extracción de la información es el mismo en todos los casos:

Se toma la distancia en segundos entre el inicio de la primera onda P y la última onda S ofreciendo una primera idea de la distancia a la que se encuentra el epicentro del terremoto del sismógrafo. Esta cifra es trasladada a la grafica izquierda y señala la distancia al epicentro, en este caso, continuando con el análisis, a 215km, y se traslada al gráfico derecho y señalando un punto. Sobre este punto se coloca una regla atravesando los puntos del grafico identificados como la distancia al epicentro y la amplitud. El punto donde la regla corte la línea media en el grafico marca la magnitud del terremoto, en este caso de 5.0.

Para identificar el epicentro se empleará la técnica de la triangulación y se precisará de un compás, un mapa y todos los registros del terremoto disponible. Tras identificar la escala del mapa, se coloca la punta del compás sobre la localización de cada uno de los sismógrafos cuyos sismogramas alertando de un mismo temblor se han estudiado y se trazará una circunferencia cuyo radio será la distancia a la que fue detectado de

[10] http://www.geo.mtu.edu/UPSeis/locating.html

cada sismógrafo en la escala del mapa. El epicentro será el punto de donde intersecten todos los radios.

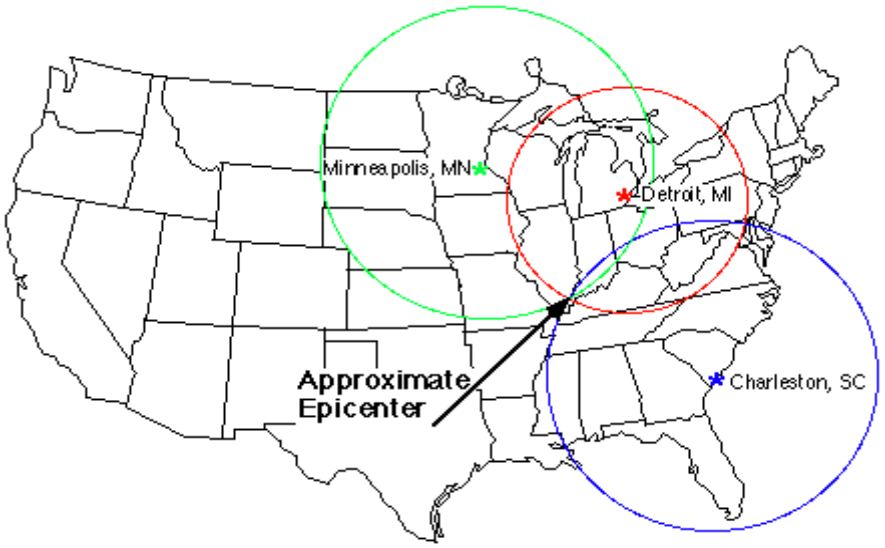


FIGURA 9 - Detección del epicentro de un terremoto empleando la triangulación[11]

2.2.6 Clasificación de señales volcánicas

La información de este epígrafe fue facilitada en su mayoría por el Instituto de Geología y Minas de Colombia quiénes me cedieron muy amablemente transparencias que ellos mismos empleaban para dar clase a los estudiantes y me ayudaron a comprender el dominio en profundidad.

La sismicidad volcánica es ocasionada por fracturas de sedimentos sólidos (rocas) movimiento y tránsito de fluidos fruto de la actividad en el interior de los conductos de un complejo volcánico.

Estos eventos se perciben en la superficie por la ocurrencia de ondas sísmicas, un tipo de ondas mecánicas, perturbaciones que se desplazan a través de un medio elástico, es decir, deformable y en el que se propagan dos magnitudes físicas conocidas como momento (cantidad de movimiento) y energía.

La energía de estas ondas se disipa en forma de calor, de pérdidas en el medio o en las fronteras del medio hasta su completa extinción.

Las señales sísmicas volcánicas se pueden clasificar de acuerdo a muchos parámetros, atendiendo a su forma geométrica (identificando las ondas como planas, cilíndricas o esféricas), el medio por el que se propagan clasificándose como internas (longitudinales o transversales) o externas (ondas Love y ondas Rayleigh).

Otro parámetro habitual es el origen o fuente originadora de la señal. En este caso se reconocen cuatro tipos (véase la Tabla 1):

[11] http://www.geo.mtu.edu/UPSeis/locating.html

Tipo de Sismo	Causa	Frecuencias dominantes
Volcano Tectónico, VT	Fractura de rocas adyacentes a conductos o depósitos de magma	5-15 Hz
Largo Período, LP	Interacción de gas o fluidos en el interior de los conductos volcánicos	Inferiores a 5 Hz
Tremor Volcánico, TR	Asociado a erupciones volcánicas	Inferiores a 5 Hz
Sismos de Hielo, HI	Origen no establecido. Posible relación con la apertura de grietas o fracturas en glaciares	Variable

Tabla 1 - Sismos en base a su origen

Otras señales de origen no volcánico que se pueden registrar son las siguientes. En la tabla que se adjunta se ordenan por el intervalo que separa la llegada de las ondas P y de la S:

Tipo de Sismo	Intervalo entre ondas S y P
Vulcano-Tectónico, VT	S-P ≤ 5 s
Tectónico Local, TL	5s < S-P ≤ 15s.
Regional, RE	15s < S-P ≤ 60s
Tele-sismo ó sismo distante, DS	S-P > 60s.

Tabla 2 - Tipos de Sismos en función del intervalo S-P

Los arribos de las ondas S y P permiten establecer el valor de la polaridad, es decir, la dirección del primer impulso de P, si fue hacia arriba o compresiva (U/D) o hacia abajo o distensiva (D) y cómo se liberó la energía, si de una manera súbita o gradual catalogándose como impulsivos o emergentes.

A continuación, se detallan los distintos tipos de ondas sísmicas y su representación visual facilitadas por las imágenes cedidas por el Instituto de Geología y Minas de Colombia (véanse las Fig. 12-21):

Sismos Volcano Tectónicos (VT): se generan por la fractura de rocas adyacentes a depósitos o conductos de magma. Tienen una forma en su envolvente parecida a un triángulo. Las ondas P y S se diferencian claramente. Tiene frecuencias entre los 5 y los 15 Hz (en ocasiones, mayores). Pueden ocurrir en paquetes o trenes (conjunto de ondas) seguidos denominados enjambres. Véanse sismograma (Figura 10) y rango de frecuencias (Figura 11):

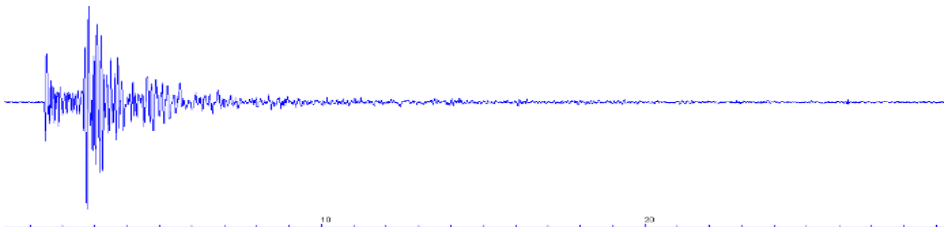


FIGURA 10 - Sismograma de un sismo de tipo VT[12]

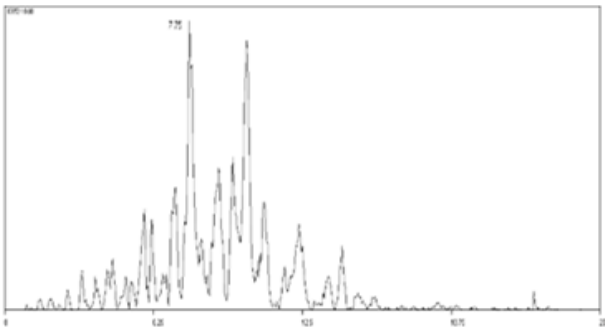


FIGURA 11 – Rango de frecuencias en un sismo VT

Sismos de Largo Período (LP): se generan por la interacción de gas, agua, fluidos hidrotermales o magma dentro de los conductos volcánicos. Tienen una forma alargada, y no es posible diferenciar las ondas S, por lo que es difícil localizar su hipocentro empleando los métodos convencionales. Poseen por lo general, frecuencias bajas inferiores a 5Hz. Véanse sismograma (Figura 12) y rango de frecuencias (Figura 13):

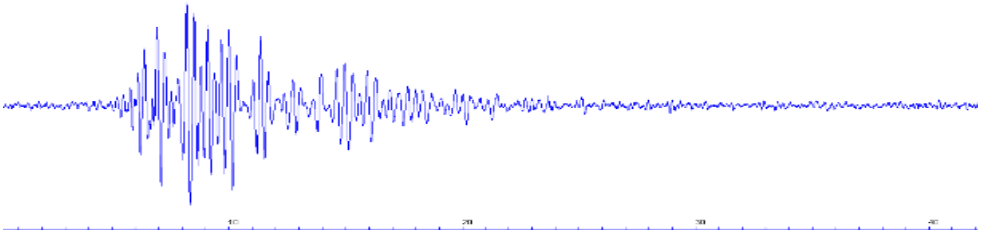


FIGURA 12 - Sismograma de un sismo de tipo LP

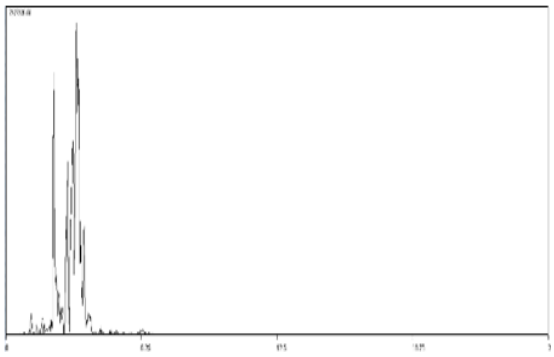


FIGURA 13 – Rango de frecuencias en un sismo LP

[12] Figuras 10 - 21 cedidas por el Instituto de Geología y Minas de Colombia.

Tremor volcánico (TR): señal sísmica cuyo origen, forma y frecuencias son similares a los de los sismos LP. Ocurren en forma de pulsos señales de corta duración, de unos pocos minutos, o bien en forma de bandas que se repiten durante horas o días continuos intermitentemente. Véanse sismograma (Figura 14) y rango de frecuencias (Figura 15):

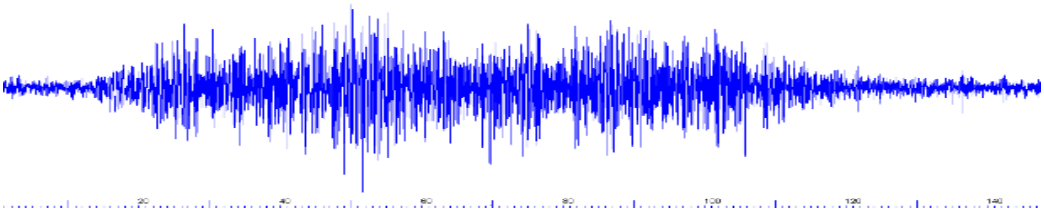


FIGURA 14 - Sismograma de un sismo de tipo TR

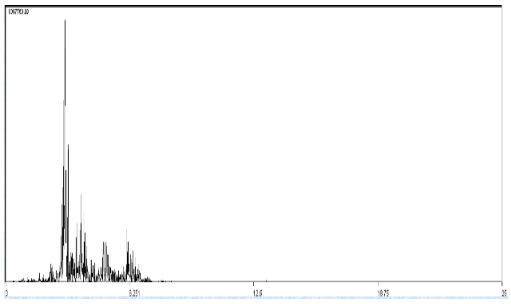


FIGURA 15 - Rango de frecuencias en un sismo TR

Sismos Híbridos (HB): señales sísmicas que representan fenómenos relacionados con las fracturas de rocas y con la dinámica de fluidos, distinguiéndose en su espectro de frecuencias una banda relacionada con la componente de fluidos y otra con el fenómeno de fractura. Véanse sismograma (Figura 16) y rango de frecuencias (Figura 17):

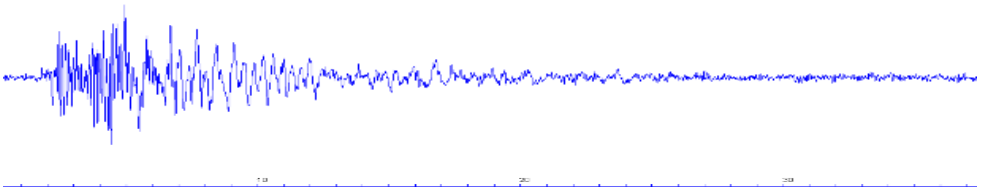


FIGURA 16 - Sismograma de un sismo de tipo HB

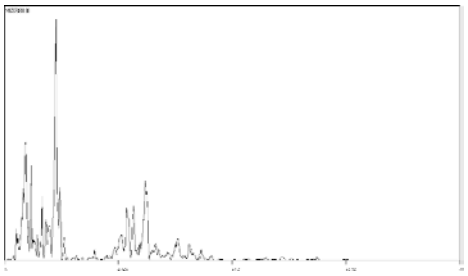


FIGURA 17 - Rango de frecuencias de un sismo HB

Sismos de tipo Tornillo (TO): eventos sísmicos de largo periodo (LP) que se caracterizan por presentar un decaimiento suave a través del tiempo y por tener espectros de frecuencia monocromáticos. Véanse sismograma (Figura 18) y rango de frecuencias (Figura 19):

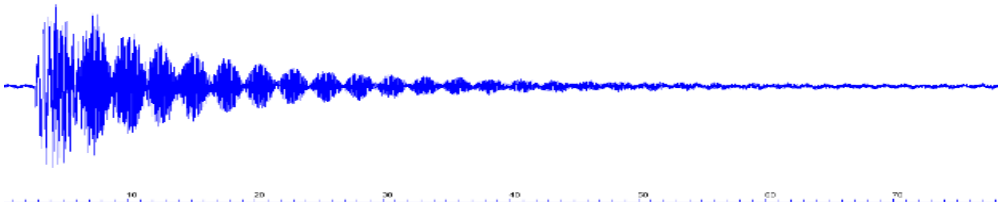


FIGURA 18 - Sismograma de un sismo de tipo TO

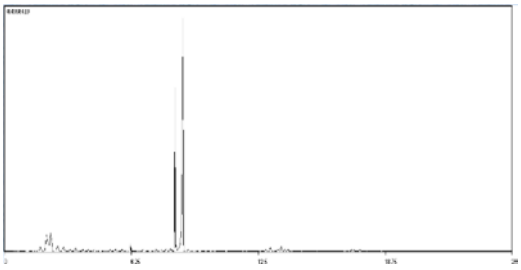


FIGURA 19 - Rango de frecuencias de un sismo TO

Actividad superficial (SU): señales sísmicas que corresponden a fenómenos que suceden a nivel superficial, tales como avalanchas, hielos, rayos, etc. Se caracterizan por presentar espectros de frecuencia ricos en altas frecuencias, indicando así, un evento asociado a la dinámica de fluidos. Véanse sismograma (Figura 20) y rango de frecuencias (Figura 21):

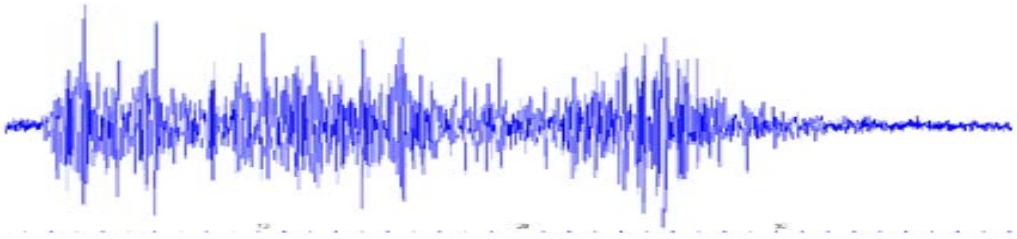


FIGURA 20 - Sismograma de un sismo de tipo SU

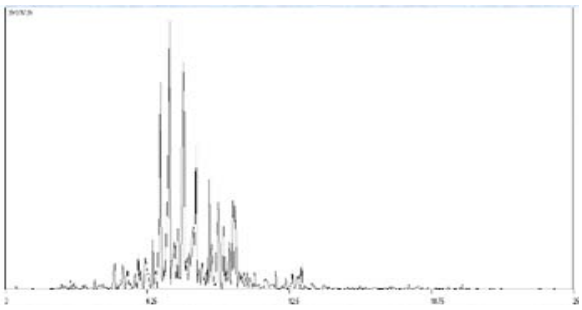


FIGURA 21 - Rango de frecuencias de un sismo SU

Tectónico local (TL): sismos originados por **fallas activas cercanas** al edificio volcánico. Entre las ondas S-P pueden pasar entre 5 y 20 segundos.

Regional (RE): sismos originados muy **lejos del edificio volcánico**, pueden tener un intervalo de ondas S-P de 50 ó 60 segundos.

Distantes (DS): sismos tectónicos originados en **otros continentes o países**.

Otros sismos, no incluyen descripción por tener un nombre suficientemente descriptivo:

Sismo por explosión en mina (EM)

No clasificable (NC)

Calibración (CA)

Rayo (RY)

Explosión (EX)

Avalancha (AV)

Baja frecuencia (BF)

No determinado (ND)

Hielo (HI)

3. Minería de datos

3.1 Introducción

La actividad sísmica de las últimas dos décadas ha sido devastadora provocando miles de víctimas y haciendo de la sismología un área de investigación muy interesante para la comunidad científica.

La sismología era hasta hace unos pocos años una ciencia accesible desde disciplinas como la física o la geología y sin embargo gracias al auge del tratamiento y análisis de datos también de la informática.

El soporte actual sobre el que se almacenan datos ya no es el papel sino el ordenador y el lenguaje no es el propio de cada país sino el binario. A estos avances, se suma el aumento del volumen de datos que actualmente se pueden almacenar y tratar de forma analítica. Todos estos avances hacen posible el estudio de esta disciplina desde el punto de vista del *Descubrimiento de Conocimiento en Bases de Datos* o KDD (por sus siglas en inglés).

Estos cambios en la forma de almacenar la información junto al auge de la Inteligencia Artificial a partir de los años 60 del siglo pasado fueron observados por ingenieros en una gran variedad de disciplinas, especialmente en la informática y aún sin saber el alcance de estos cambios, Gregory Piatetsky-Shapiro vislumbró todas sus posibilidades, concretamente, la identificación de patrones y de información oculta en grandes repositorios que eran susceptibles de extraerse empleando técnicas de Aprendizaje Automático.

En el taller que ofreció en 1989 (Piatetsky-Shapiro January 1991) bautizó como KDD (*Knowledge Discovery from Databases*) a este fenómeno y ofreció la siguiente definición para KDD:

"proceso no trivial de identificar patrones válidos, novedosos, potencialmente útiles y en última instancia comprensibles a partir de los datos." (Fayyad U. M 1996)

Este término fue adoptado por la comunidad científica, pero otro término se hizo mucho más popular entre la población: data mining o en español, minería de datos adoptado por empresas e investigadores en bases de datos en lugar de KDD, al igual que se hará en este Trabajo de Fin de Grado (Mena 2011).

3.2 Minería de Datos

3.2.1 Aplicaciones

Dado su carácter interdisciplinar, la minería de datos se emplea en muy distintos ámbitos para dar solución o aportar información a distintas problemáticas siempre y cuando cumplan con los siguientes supuestos: disponer de un gran poder de cómputo y contar con una gran cantidad de datos.

A continuación, un listado representativo de algunos de estos dominios acompañados de ejemplos:

- **Atención al cliente:** identificar patrones de compra, clasificar los clientes susceptibles de encontrar atractiva una campaña.

- **Banca:** detección de patrones de uso fraudulento de tarjetas de crédito, identificar reglas de mercado de valores a partir de históricos.

- **Logística:** predecir la cantidad de producto necesaria que es necesario enviar a un determinado almacén en un determinado instante.

- **Seguros:** identificar patrones de comportamiento en clientes que puedan incurrir en fraude, predecir posibles gastos médicos de un cliente que sufre una enfermedad y está bajo tratamiento.

- **Entretenimiento:** analizar los gustos de la audiencia para establecer una determinada programación y unas determinadas cuñas de publicidad, predecir el éxito de una película para realizar una inversión adecuada en la misma.

- **Deportes:** análisis de vídeos de los jugadores del equipo contrario para descubrir su estrategia.

- **Medicina:** asociar síntomas y clasificar patologías, predecir tasas de éxito en trasplantes que ayuden a mejorar las parejas donante-receptor. (Turban 2011), (García Jiménez 2005).

3.2.2 Alcance de la Minería de Datos

El alcance de la minería de datos está estrechamente relacionado con los conocimientos del analista acerca del dominio, su experiencia, capacidad de observación y perspicacia.

Las técnicas de minería de datos arrojan resultados inesperados que es necesario interpretar. Sin interpretación, no hay respuesta y sin respuesta, persiste el problema. Cuando esta interpretación ocurre, el conocimiento del problema se amplía pudiendo llevar incluso a una solución.

En suma, las técnicas de minería de datos permiten:

- **Predecir tendencias y comportamientos:** averiguar el porqué, es lo que puede explicar porqué un determinado patógeno es inmune a un medicamento pero en combinación con otros puede ser tratado o realizar una tarea tan simple como aumentar o bajar el precio de un inmueble que está a la venta para atraer a un cliente.

- **Descubrimiento de modelos o patrones desconocidos**: a menudo se confunden causalidad y casualidad. Tendemos a pensar que si el evento B posterior a A ha ocurrido un determinado número de veces, la próxima vez que ocurra A, creeremos que ocurrirá B. Técnicas de minería de datos podrían demostrar que B ocurre por un

evento C o por A si también se da C o que la relación entre ambos eventos es débil

- **Analizar las BBDD (Bases de Datos) con todas sus filas y columnas**: eliminar columnas puede conllevar eliminar información relevante y eliminar filas incurrir en errores de estimación con valores altos. Las técnicas de minería de datos permiten realizar análisis sin reducir la dimensionalidad del conjunto de entrada. (Blázquez García 2004).

3.2.3 Análisis de Datos

El primer paso para llevar a cabo un análisis de datos es su obtención. Ésta se lleva a cabo recuperando datos de una BBDD o de otras fuentes. Los datos constituyen la unidad mínima de abstracción de la que es posible extraer información y/o conocimiento. Se obtienen por medio de la observación, la experiencia o la experimentación y se presentan en formatos muy distintos como números, palabras, imágenes, etc. (Turban 2011).

Previo al análisis, es necesario inspeccionar los datos y realizar una **limpieza** y **selección** de los mismos. En la **limpieza** se eliminan los **datos anómalos y ausentes** resultando útiles la elaboración de histogramas que permitan detectar rápidamente los datos que no siguen la distribución general. Se pueden llevar a cabo las siguientes acciones:

- **Ignorar:** esta acción es recomendable llevarla a cabo cuando los algoritmos son sensibles ante este tipo de datos.

- **Reemplazar el valor:** está técnica es recomendable si es posible reemplazar el valor atípico por otro que esté basado en el conjunto de datos que se está utilizando.

- **Discretizar:** esta acción resulta muy útil al homogeneizar los datos. Los atípicos simplemente pasan a estar agrupados en torno a un valor o a una etiqueta.

- **Eliminar o reemplazar la columna:** es la opción menos recomendable porque podríamos perder información valiosa, convendría tratar la columna con alguna de las técnicas anteriormente citadas.

- **Filtrar la fila:** alterar directamente una instancia es introducir un sesgo, pero éste podría no tener importancia si el conjunto de datos a analizar es lo suficientemente extenso y el algoritmo robusto.

- **En el caso de datos ausentes,** se podría estudiar la posibilidad de aguardar a que estén disponibles. Este caso es complejo y conviene saber el porqué de su ausencia para tomar la mejor decisión:
  - **El valor no existe:** es posible que un determinado dato no se dé si no se ha cumplido ciertas condiciones, por ejemplo, en el

tratamiento y diagnóstico de una enfermedad para saber cómo continuar con el tratamiento puede ser necesario esperar la reacción del paciente ante un fármaco.

- **La ausencia tiene significado:** el hecho de que un dato no aparezca puede ser indicativo de alguna característica relevante.

- **Información incompleta:** si los datos provienen de distintas fuentes es posible que algunos de los campos estén incompletos ya que al agrupar la información, la tendencia es unir los campos no intersecarlos. Un ejemplo, es la información detectada por una serie de sensores que se transmite a una estación de radio y está a su vez a una estación de control. En la estación de radio, las señales se juntan unas con otras, pero no se presta atención al valor de los campos.

La selección de datos se lleva a cabo con el propósito de eliminar aquellos dates irrelevantes o que pueden ser sustituidos por otro dato que guarde una relación más clara con el dato/s eliminado/s o el conjunto de datos. De acuerdo a los criterios y políticas de selección será posible:

- Realizar un muestreo vertical u horizontal eliminando datos.

- Elaborar nuevos atributos operando con los presentes en el conjunto de datos e incluyéndolos, pudiendo añadir una nueva columna y/o borrando los datos empleados para crear el nuevo atributo.

3.2.4 Aprendizaje en Minería de Datos

En la elaboración de cualquier modelo o en la detección de cualquier patrón en un conjunto de datos está presente una fase de aprendizaje, en la que es preciso estudiar el set de datos, extraer conocimiento y aprehenderlo.

El aprendizaje supervisado es aquél en que el que por cada fila (o instancia) hay una etiqueta que indica su categoría o coste y permite al algoritmo su agrupación en base a cada uno de los valores de la clase.

El aprendizaje no supervisado en cambio, no emplea etiquetas. Los datos son analizados en bruto y los patrones extraídos de los datos son intrínsecos a la información que hay en ellos.

3.2.5 Técnicas de Minería de Datos

Ante el incremento en la capacidad de almacenamiento de datos experimentada en los últimos años, la velocidad a la que se generan y la inmediatez en la respuesta que se desea obtener, se están desarrollando nuevas técnicas en la minería de datos que atiendas estos requisitos y permiten elaborar modelos de conocimiento en tiempo real con la menor cantidad de información posible almacenada en memoria. (Ferrer–Troyano 2005).

En cambio un esquema de aprendizaje incremental responde a las siguientes características:

- Integra la dimensión temporal en el análisis de los datos. En función de la tarea que se esté analizando, la dimensión temporal será un atributo del conjunto de entrenamiento o un parámetro relevante para el proceso de aprendizaje. (Ferrer–Troyano, 2005).
- Incorpora algoritmos robustos sensibles a los datos atípicos y a la temporalidad. En el aprendizaje por lotes, el conjunto de entrenamiento disponible encierra toda la realidad. Este supuesto es conocido como mundo cerrado frente a su opuesto, mundo abierto, del que parten los algoritmos incrementales. (Ferrer–Troyano, 2005). Son por lo tanto, sistemas flexibles que ante la detección de determinadas circunstancias, como un determinado número de instancias o de tiempo o datos atípicos modifican las reglas que estaban siguiendo para generar el modelo con esta nueva información.
- Curva de aprendizaje. Los sistemas por lotes generan un único patrón o modelo siendo este modelo la muestra máxima de su conocimiento. Un sistema incremental en cambio, crece y evoluciona a lo largo del tiempo lo que provoca que su fiabilidad sea mayor cuanto más tiempo haya estado ejecutándose el algoritmo.

Los sistemas de decisión off–line son los empleados tradicionalmente en Minería de Datos. Empleando un único conjunto de datos, éste era procesado iterativamente hasta elaborar un modelo o extraer un patrón de acuerdo a los siguientes tres supuestos:

- El dominio del problema está claramente descrito en los ejemplos presentes en el conjunto de datos antes del proceso de aprendizaje.
- Todos los ejemplos de entrenamiento pueden ser cargados en memoria.
- Tras procesar debidamente el conjunto de entrenamiento, el aprendizaje puede darse por finalizado.

Los sistemas de decisión on-line son aquellos capaces de dar una respuesta en todo momento y pueden seguir un esquema de aprendizaje off-line o incremental.

Tanto un problema como un esquema de aprendizaje pueden ser incrementales. Una tarea de aprendizaje es incremental cuando el conjunto de entrenamiento sobre el que se quiere aprender se va generando a lo largo del tiempo y es procesado secuencialmente obligando al sistema a aprender en episodios sucesivos (Ferrer–Troyano 2005).

Los sistemas de decisión on-line pueden funcionar por lotes o desde un punto de vista incremental. Un sistema de decisión on-line que trabaje por aprendizaje por puramente incremental tan sólo tiene en cuenta los ejemplos recibidos en cada instante y es posible obtener buenos resultados.

Un sistema de decisión on-line que trabaje por aprendizaje por lotes sólo arrojará resultados satisfactorios si es capaz de cargar todos los datos en memoria y tratar el problema desde el punto de vista del aprendizaje incremental. Sería necesario por tanto que desechara y generara un nuevo modelo de aprendizaje por cada nuevo conjunto de entrenamiento formado por la adición de una nueva instancia al conjunto de entrenamiento anterior cada vez que ésta se produzca.

Este esquema se conoce como – *temporal–batch* – y tiene como única ventaja respecto al aprendizaje por lotes tradicional que el orden de llegada de los ejemplos no afecta al modelo creado. (Ferrer–Troyano 2005). Sin embargo, son ineficientes y no ofrecen buenos resultados cuándo son probados en dominios dinámicos reales al no estar preparados los algoritmos de aprendizaje para integrar la nueva información aportada por nuevas instancias al siguiente modelo.

3.2.6 Patrones detectados en la Minería de Datos

La minería de datos permite construir modelos matemáticos empleando los sistemas de decisión online y off-line con esquemas de aprendizaje incrementales y no incrementales. Estos sistemas tratarán de identificar uno de los siguientes cuatro tipo de patrones elaborando un modelo de predicción, asociación o segmentación:

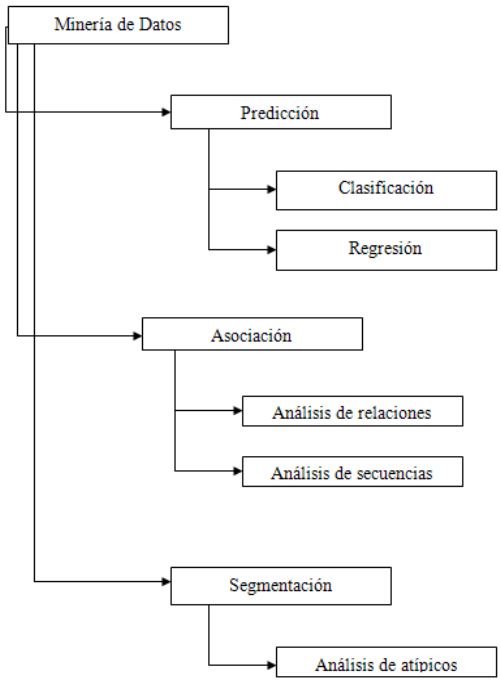


FIGURA 22 - Taxonomía de tareas en Minería de Datos [basada en CRISP-DM, 2000]

3.2.6.1 Predicción

Basándose en la información extraída en sucesos pasados explica el valor o predice la ocurrencia de un determinado suceso. Esta predicción se puede llevar a cabo sobre un número de categorías específicas (clasificación) o sobre determinados valores numéricos (regresión). Las técnicas que se empleen puede además ser o no ser

explicativas y han de ser elegidas en función de si se persigue una mayor comprensión del dominio (árboles de decisión) o simplemente resultados (redes de neuronas).

3.2.6.2 Asociación

Las reglas de asociación se utilizan para encontrar hechos que ocurren de manera simultánea de forma habitual en un conjunto de datos. Es decir, se detectan las condiciones bajo las que ocurren ciertos eventos.

La eficacia de un algoritmo de asociación se mide por los valores de su cobertura y precisión. La cobertura o soporte es el número de instancias predichas correctamente empleando las reglas de asociación generadas y la precisión o confianza, la proporción de número de instancias respecto al total de instancias del conjunto sobre la que es aplicada la regla.

Un algoritmo de asociación típico es A priori.

3.2.6.3 Segmentación

Aproxima elementos y los introduce en un mismo grupo basándose en su semejanza entre ellos mismos y sus diferencias con el resto. Su estudio permite establecer jerarquías y obtener nueva información (relación entre atributos, detección de datos atípicos, etc.). Las dos técnicas de agrupación más populares son los mapas de Kohonen y el algoritmo K-Medias. (Turban 2011), (Blázquez García 2004).

3.3 Herramienta de minería de datos

En este proyecto se empleará la herramienta de minería de datos Weka. Weka es una herramienta de software libre diseñada y creada por la Universidad de Waikato (Nueva Zelanda) para resolver tareas empleando algoritmos de Aprendizaje Automático.

Al tratarse de una herramienta de software libre portable y haber sido manejada durante el grado la consideré adecuada para llevar a cabo este proyecto frente a otras como Hadoop que hubieran requerido de aprendizaje por mi parte.

Con esta herramienta se estudiarán las dependencias de los atributos del dominio y generarán modelos de clasificación supervisada y no supervisada con técnicas basadas en aprendizaje incremental y aprendizaje por lotes. WEKA puede ser utilizada desde línea de comandos o desde su propia GUI abundando información de acceso libre para saber cómo hacerlo. La versión empleada en este proyecto es la **versión 3.9.1** a la que se añadieron de manera externa paquetes que aumentaban su funcionalidad. Estos paquetes permiten realizar aprendizaje incremental en lugar de únicamente aprendizaje por lotes. Este enfoque permite a un usuario experto en WEKA poder manejar algoritmos de aprendizaje automático desde un punto de vista incremental sin tener que aprender a manejar otra herramienta.

3.4 Metodología CRISP-DM

En este Trabajo de Fin de Grado se seguirá la metodología **CRISP-DM** (*Cross Industry Standard Process for Data Mining*). CRISP-DM es un modelo de proceso de minería de datos compuesto por un modelo (véase la Fig. 23) y una guía cuyas fases se detallan a continuación (Rojas 2010).[13][14]:

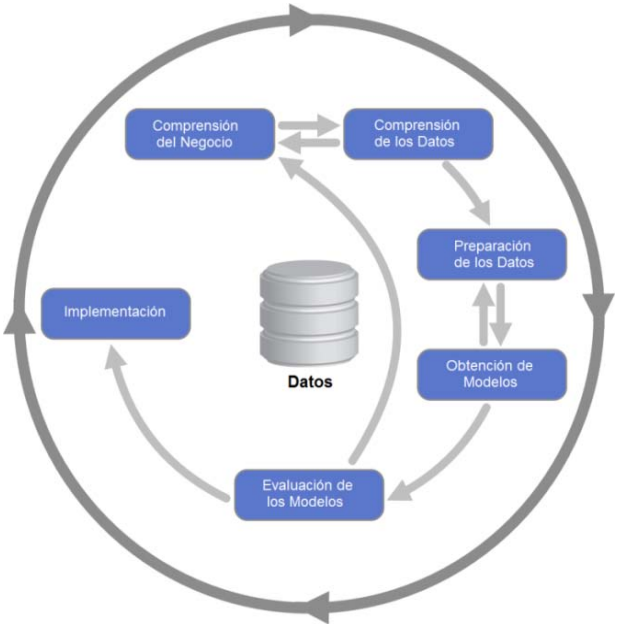


FIGURA 23 - CRISP-DM

3.4.1 Comprensión del negocio.

Esta primera fase los objetivos y requisitos del proyecto desde una perspectiva de negocio se traducen a objetivos técnicos y a un plan de proyecto.

Es indispensable establecer estos objetivos y comprenderlos para elaborar algoritmos cuya funcionalidad se corresponda con la exigida en el mundo real, sin ello, no será posible recolectar los datos correctos e interpretar correctamente los resultados.

Resulta imprescindible, la capacidad de convertir el conocimiento adquirido del negocio en un problema de minería de datos y en un plan preliminar que permita alcanzar los objetivos del negocio. Véase la Fig. 24 para una descripción de cada una de las tareas que componen esta fase:

[13] Todas las imágenes que se adjuntan se podrán encontrar en [CRISP-DM, 2000]

[14] http://bizmetriks.com/metodologia.html

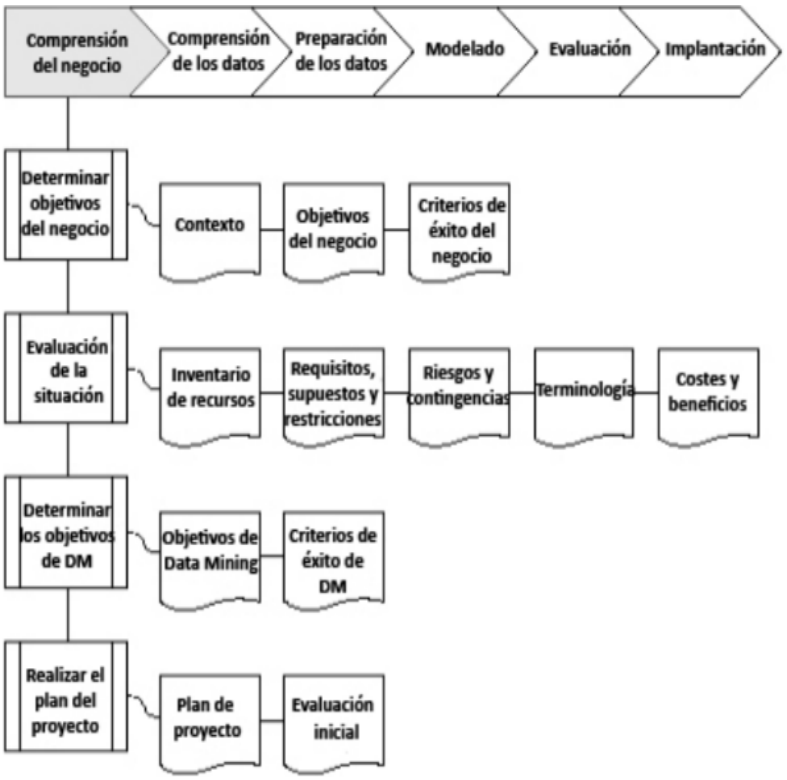


FIGURA 24 – Fase de comprensión del negocio

3.4.2 Comprensión de los datos.

Esta segunda fase comprende la recolección inicial de los datos con el objetivo de establecer un primer contacto con el problema, familiarizarse con ellos, identificar su calidad y establecer las relaciones más evidentes que permitan definir las primeras hipótesis. Esta fase junto a las dos siguientes fases son las que demandan el mayor esfuerzo y tiempo en un proyecto de minería de datos.

Por lo general si la organización cuenta con una base de datos corporativa, es deseable crear una nueva base de datos específica para el proyecto de minería de datos, ya que durante el desarrollo del proyecto es posible que se generen frecuentes y abundantes accesos a la base de datos con el fin de realizar consultas y probablemente se produzcan modificaciones, lo cual podría generar muchos problemas. Véase la Fig. 25:

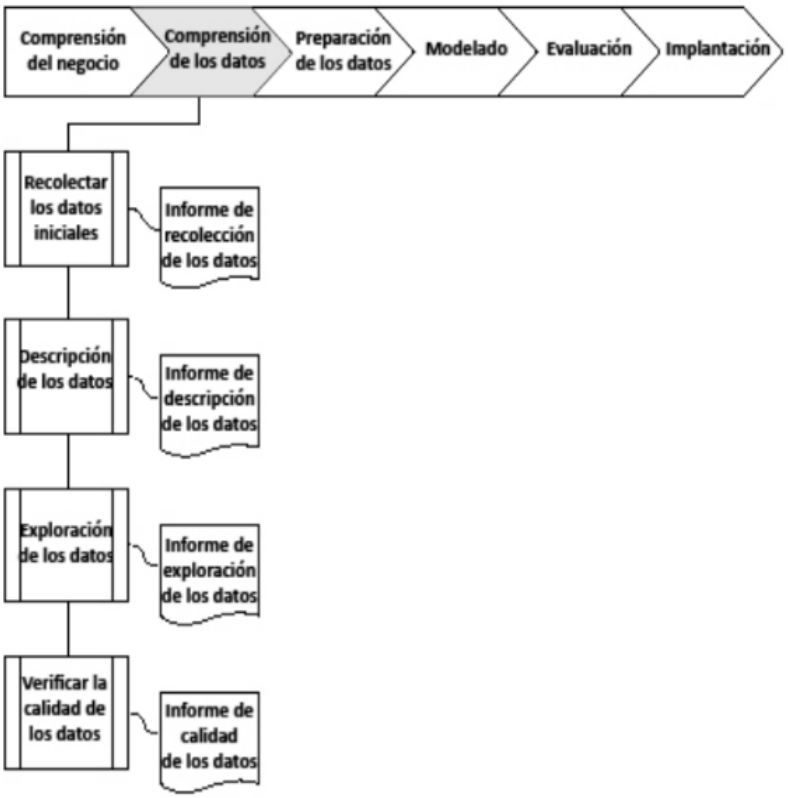


FIGURA 25 - Fase de comprensión de los datos

3.4.3 Preparación de los datos.

En esta fase, se preparan los datos para adaptarlos a las técnicas de minería de datos que se van a utilizar (técnicas de visualización de datos, búsqueda de relaciones entre variables u otras medidas para explotación de los datos).

La preparación de los datos incluye las tareas generales de selección de datos a los que se va a aplicar una determinada técnica de modelado, limpieza de datos, generación de variables adicionales, integración de diferentes orígenes de datos y cambios de formato.

Esta fase se relaciona con la fase de modelado. En función de la técnica elegida, los datos se procesarán de una manera u otra, por esta razón las fases de preparación y modelado interactúan de forma permanente y se retroalimentan.

Véase la Fig.26 para conocer las tareas que se llevan a cabo en esta fase:

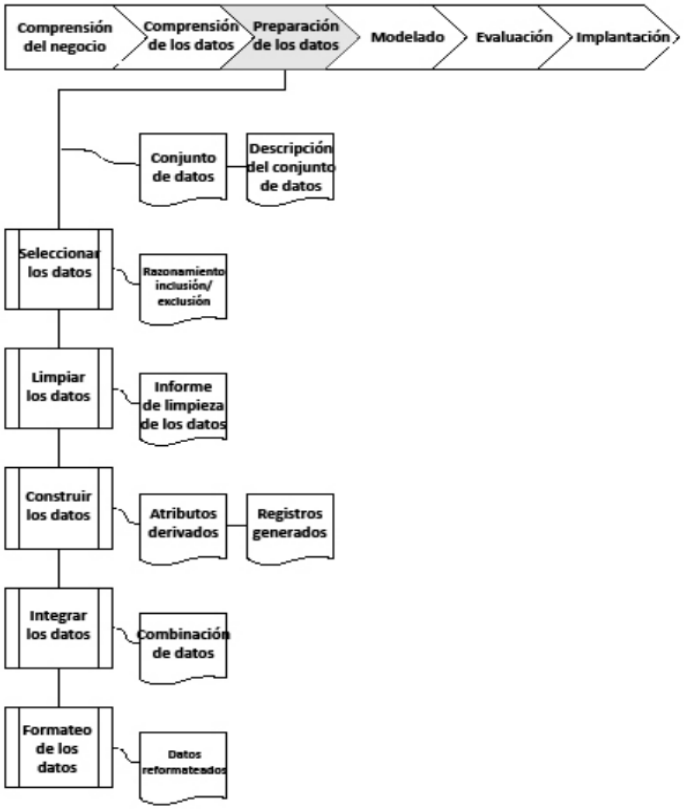


FIGURA 26 - Fase de preparación de los datos

3.4.4 Modelado u Obtención de Modelos.

En esta fase de CRISP-DM se seleccionan las técnicas de modelado más apropiadas para el proyecto de minería de datos que se esté tratando en función de los siguientes criterios:

- Son apropiadas para el problema.
- Los datos son adecuados.
- La técnica cumple con los requisitos del problema.
- Genera un modelo en un intervalo de tiempo adecuado.
- Se conoce la técnica y cómo interpretarla.

Se pueden elegir una o varias técnicas así como métodos de evaluación que permitan establecer su grado de adecuación. Los parámetros que se utilicen en la generación del modelo dependerán de las características de los datos y de las características de precisión que se quieran lograr con el modelo. Véanse las tareas que se llevan a cabo en esta fase en la Fig.27:

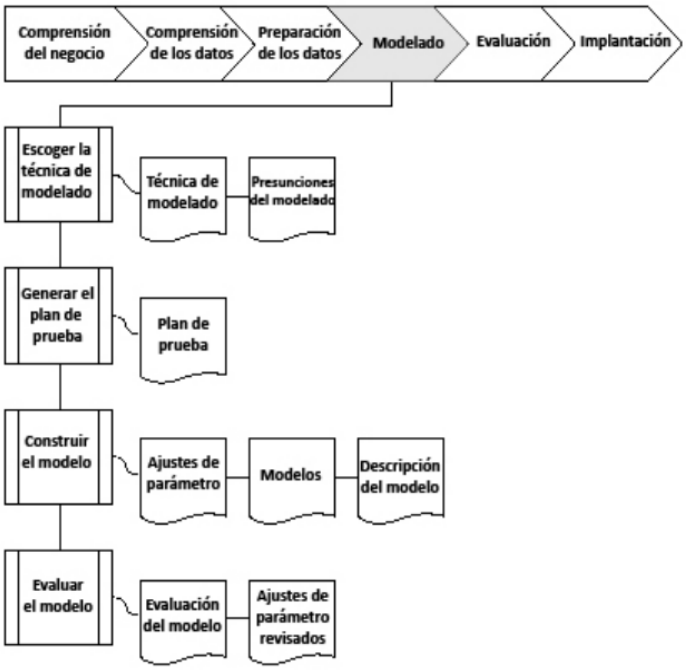


FIGURA 27 - Fase de modelado

3.4.5 Evaluación.

En esta fase se evalúa el modelo, teniendo en cuenta el cumplimiento de los criterios de éxito del problema. Debe considerarse además que la fiabilidad calculada para el modelo se aplica solamente para los datos sobre los que se realizó el análisis. Es preciso revisar el proceso, teniendo en cuenta los resultados obtenidos, para poder repetir algún paso anterior, en el que se pueda haber cometido algún error. Considerar que se pueden emplear múltiples herramientas para la interpretación de los resultados. Si el modelo generado es válido en función de los criterios de éxito establecidos en la fase anterior, se procede a la explotación del modelo.

Las tareas involucradas en esta fase del proceso se detallan en la Fig. 28:

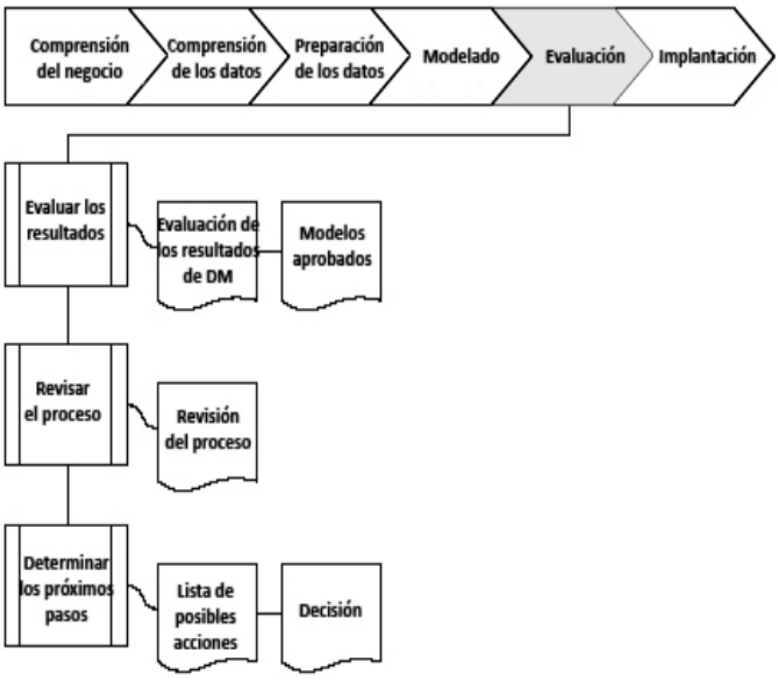


FIGURA 28 - Fase de evaluación

3.4.6 Despliegue o implantación.

En esta fase, y una vez que el modelo ha sido construido y validado, se transforma el conocimiento obtenido en acciones dentro del proceso de negocio, esto puede hacerse por ejemplo cuando el analista recomienda acciones basadas en la observación del modelo y sus resultados, o por ejemplo aplicando el modelo a diferentes conjuntos de datos o como parte del proceso (en análisis de riesgo de créditos, detección de fraudes, etc.).

Generalmente un proyecto de minería de datos no concluye en la implantación del modelo, ya que se deben documentar y presentar los resultados de manera comprensible para el usuario con el objetivo de lograr un incremento del conocimiento.

Por otra parte, en la fase de explotación se debe asegurar el mantenimiento de la aplicación y la posible difusión de los resultados. Véase la Fig. 29 para conocer las tareas que componen esta fase:

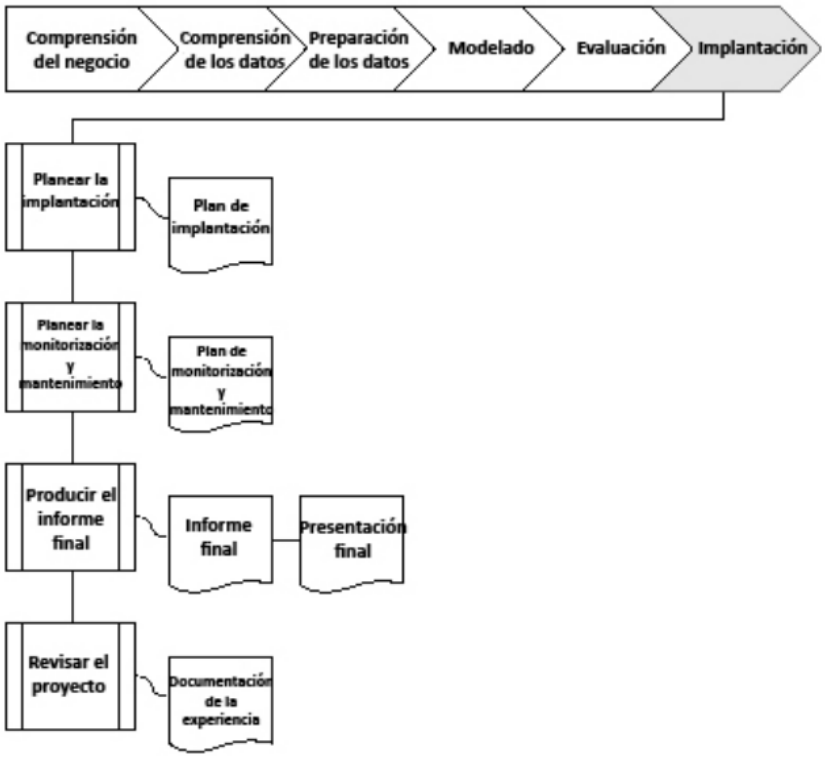


FIGURA 29 - Fase de implantación

Este Trabajo de Fin de Grado no aborda la fase de implantación al estar fuera del alcance del proyecto. En otras palabras, se llevan a cabo sólo las 5 primeras fases de la metodología CRISP-DM.

4. Desarrollo de la propuesta

4.1 Comprensión del Negocio

Las técnicas de minería de datos son cada vez más populares en el campo de la sismología observándose dos líneas de trabajo bien diferencias entre sí: la repetición de experimentos con técnicas de minería de datos llevados a cabo hace 20 años (McNutt 1992) (Mcnutt 1994) (Unglert K. 2017) y la predicción de terremotos (National Institute of Geophysics and Volcanology 2016).

En el pasado los conjuntos de datos con los que era posible trabajar debían de no superar un determinado tamaño y cuidar que los datos fueran relevantes, sin apenas ruido (Mcnutt 1994) para llevar a cabo análisis fiables.

Sin embargo, estas herramientas no se emplean todavía en tareas como la localización de epicentros o la interpretación de los datos de inclinómetros (Alpala J. 2017). ¿Podría el uso de tareas de minería de datos realizar de manera más fiables estas tareas? Es más, ¿podrían arrojar información oculta que sirviera para elaborar técnicas más robustas dentro del campo de la sismología como es la detección de metales pesados? (Telenchana E. 2017)

Este trabajo se centrará en realizar modelos de clasificación que permitan clasificar las ondas detectadas en un sismograma de forma no manual y despertar el interés dentro de la comunidad científica de uso de técnicas de minería de datos en el área de la sismología.

4.1.1 Evaluación de la situación

Los recursos que se emplearán en este trabajo serán el entorno de trabajo Eclipse, Notepad++, el paquete de Microsoft Office, la herramienta de minería de datos Weka y los registros de la monitorización volcánica del volcán Puracé en el período comprendido entre el 01-07-2015 y el 01-07-2016.

El objetivo de este trabajo es la elaboración de modelos de clasificación de los 18 tipos de ondas sísmicas que ocurren en torno al volcán Puracé. La elaboración de los clasificadores incluyó la realización de varias tareas: análisis exploratorios sobre los datos, elaboración y evaluación de los modelos de clasificación obtenidos y experimentos de segmentación que respaldaran los resultados.

4.2 Comprensión de los Datos

La tarea de recabar los datos fue llevada a cabo por el grupo de Control, Aprendizaje y Observación de Sistemas (CAOS) de la Universidad Carlos III de Madrid. Este grupo contactó con el Instituto Colombiano de Ingeniería y Minas por lo que la fiabilidad y veracidad de los datos está constatada. Estos ficheros se encuentran alojados en *http://www.caos.inf.uc3m.es/datasets/* y contienen la monitorización volcánica del volcán Puracé desde el año 01-07-2015 al 01-07-2016.

El repositorio Purace1Year se estructura en tres ítems, dos carpetas con el nombre del año en el que almacenan los datos y análisis de la actividad sísmica de ese período y un fichero README[15].

FIGURA 30 - Distribución Repositorio

La carpeta 2015 contiene carpetas con los meses Julio, Agosto, Septiembre, Octubre, Noviembre y Diciembre de ese año. En las carpetas en las que se incluye la terminación "_data" se almacenan los registros de los distintos sismógrafos colocadas en torno al volcán Puracé. En las carpetas en las que la terminación es "_class" el análisis correspondiente sobre los ficheros de ese mes.

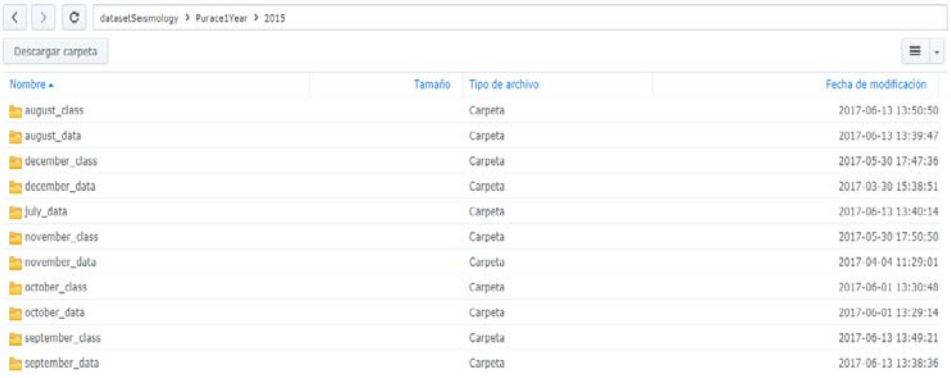


FIGURA 31 - Contenido Año 2015

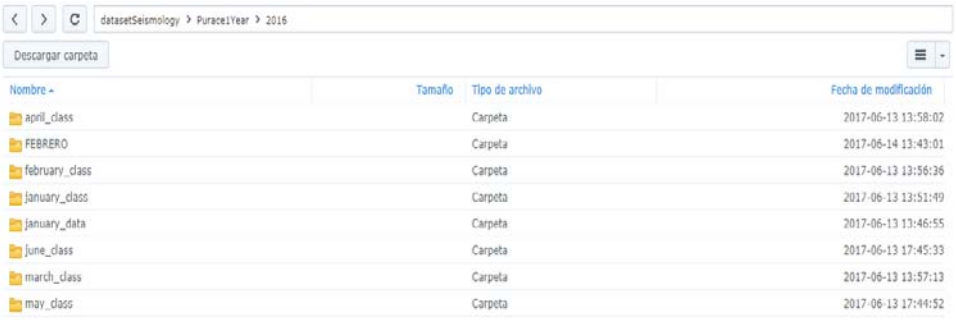


FIGURA 32 - Contenido Año 2016

Al no disponer de carpetas con la monitorización de los meses del año 2016 se resolvió llevar a cabo la experimentación con el año 2015. Las carpetas del año 2015

[15] Este fichero se adjunto pocos días antes de la entrega final de este proyecto y en el que se disculpan por no haberlo tenido accesible durante los meses de Agosto y parte de Septiembre.

tienen la siguiente estructura según acaben en _data o en _class independientemente del mes al que hagan referencia, almacenando tantos ficheros como días tenga el mes:

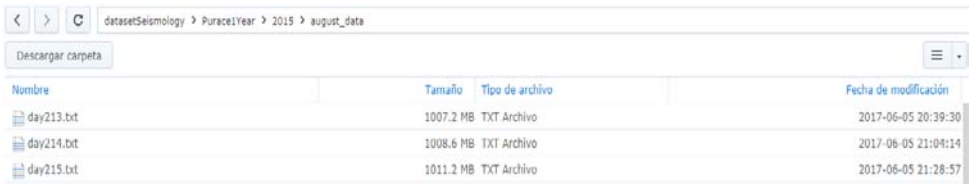


FIGURA 33 - Ficheros de la carpeta august_data

FIGURA 34 - Ficheros de la carpeta august_class

Como se puede observar, la monitorización volcánica se lleva a cabo diariamente. Cada día es una carpeta en cuyo interior hay un .txt con los datos almacenados en crudo almacenados y un fichero .csv con su contenido analizado.

4.2.1 Fichero .txt

Los ficheros .txt tienen 19 atributos en los que el carácter "?" representa un valor ausente. Tienen un tamaño de en torno a **1GB** y recogen la actividad símica diaria registrada por los sismógrafos colocados en torno al volcán Puracé. Veáse la Fig. 36 en la que se aprecia una captura de pantalla con la actividad sísmica del día 244 (day244.txt):

```
1 data,ABLO.x,ABLO.y,ABLO,COZO.x,COZO.y,COZO,COBO.x,COBO.y,COBO,LARO.x,LARO.y,LARO,SHAO.x,SHAO.y,SHAO
2 2015-08-12T00:00:00.000005,-9634,-10038,-9929,-9352,-9040,-9279,-9088,-6225,-7666,-10137,-8948,-13891,14786,-9394,-33434
3 2015-08-12T00:00:00.010005,-9648,-10044,-9921,-9422,-9084,-9179,-8922,-6449,-7628,-10126,-8940,-13894,14948,-9382,-33448
4 2015-08-12T00:00:00.020005,-9650,-10097,-9944,-9324,-9103,-9270,-9225,-6412,-7725,-10119,-8919,-13891,14964,-9109,-33297
5 2015-08-12T00:00:00.030005,-9669,-10065,-9932,-9335,-9249,-9324,-8941,-6324,-7722,-10113,-8918,-13881,14973,-8958,-33317
6 2015-08-12T00:00:00.040005,-9671,-10025,-9920,-9367,-9388,-9315,-8817,-6484,-7715,-10110,-8923,-13876,14969,-9110,-33245
7 2015-08-12T00:00:00.050005,-9667,-10081,-9948,-9405,-9063,-9269,-8737,-6334,-7662,-10125,-8913,-13869,14988,-9569,-33468
8 2015-08-12T00:00:00.060005,-9637,-10061,-9925,-9409,-9007,-9315,-8809,-6407,-7706,-10127,-8917,-13871,14983,-9578,-33566
9 2015-08-12T00:00:00.070005,-9599,-10074,-9920,-9337,-9316,-9312,-8877,-6542,-7794,-10143,-8920,-13875,15135,-8965,-33655
10 2015-08-12T00:00:00.080005,-9615,-10032,-9902,-9330,-9260,-9307,-8862,-6387,-7696,-10144,-8929,-13872,15112,-8542,-33686
11 2015-08-12T00:00:00.090005,-9579,-10027,-9905,-9456,-9092,-9313,-9025,-6420,-7725,-10159,-8935,-13861,15182,-8462,-33419
12 2015-08-12T00:00:00.100005,-9593,-10096,-9937,-9335,-9254,-9248,-9898,-6558,-7796,-10144,-8946,-13884,14946,-8748,-33223
13 2015-08-12T00:00:00.110005,-9591,-10129,-9924,-9408,-9161,-9284,-9153,-6421,-7643,-10147,-8962,-13885,14865,-9069,-33060
14 2015-08-12T00:00:00.120005,-9579,-10116,-9914,-9578,-9081,-9288,-8964,-6447,-7572,-10152,-8972,-13881,14764,-9195,-33226
15 2015-08-12T00:00:00.130005,-9594,-10102,-9903,-9309,-9285,-9312,-9078,-6420,-7547,-10160,-8992,-13887,14752,-8961,-33358
16 2015-08-12T00:00:00.140005,-9621,-10119,-9938,-9416,-9255,-9329,-8803,-6327,-7698,-10171,-9001,-13878,15016,-9037,-33605
17 2015-08-12T00:00:00.150005,-9622,-10100,-9915,-9527,-9155,-9279,-8985,-6234,-7542,-10176,-9010,-13875,15218,-9064,-33403
18 2015-08-12T00:00:00.160005,-9607,-10087,-9916,-9410,-9094,-9232,-9121,-6486,-7899,-10181,-9011,-13864,15368,-9166,-33360
19 2015-08-12T00:00:00.170005,-9627,-10030,-9952,-9345,-9046,-9260,-8979,-6462,-7660,-10163,-9028,-13891,15192,-9350,-33556
20 2015-08-12T00:00:00.180005,-9601,-10036,-9930,-9284,-9305,-9293,-8873,-6404,-7657,-10171,-9034,-13888,15061,-9353,-33573
21 2015-08-12T00:00:00.190005,-9630,-10018,-9918,-9446,-9323,-9320,-9178,-6517,-7810,-10154,-9037,-13884,14821,-9288,-33642
22 2015-08-12T00:00:00.200005,-9640,-9985,-9901,-9416,-9222,-9307,-9021,-6348,-7660,-10163,-9048,-13866,15038,-8879,-33701
```

FIGURA 35 - Ejemplo de .txt

El primer atributo que se observa es la fecha indicando año, mes, día y hora en que se registró la onda. El resto de atributos son valores numéricos que registran el cambio de voltaje de la onda sísmica sobre las componentes E, N, Z del sismógrafo al que hagan referencia (ABL, COC, CON, LAR, PIL, SHA).

Atributo		Formato
Fecha (día y hora)		Fecha
Agua Blanca (ABL)	E	Numérico
	N	Numérico
	Z	Numérico
Coucy (COC)	E	Numérico
	N	Numérico
	Z	Numérico
Condor (CON)	E	Numérico
	N	Numérico
	Z	Numérico
Lavas Rojas (LAR)	E	Numérico
	N	Numérico
	Z	Numérico
Pilimbala (PIL)	E	Numérico
	N	Numérico
	Z	Numérico
Shaka (SHA)	E	Numérico
	N	Numérico
	Z	Numérico

Tabla 3 – Atributos .txt

4.2.2 Fichero .csv

Un fichero .csv tendrá los resultados de los análisis llevados a cabo sobre los datos de un único .txt de un único día. Un .csv analizará los datos de un fichero .txt que tenga el mismo número que él. Es decir, el análisis del fichero .txt, day244.txt se encuentra en el fichero .csv, 244.csv (11 de Agosto de 2015). Véase la Fig. 37:

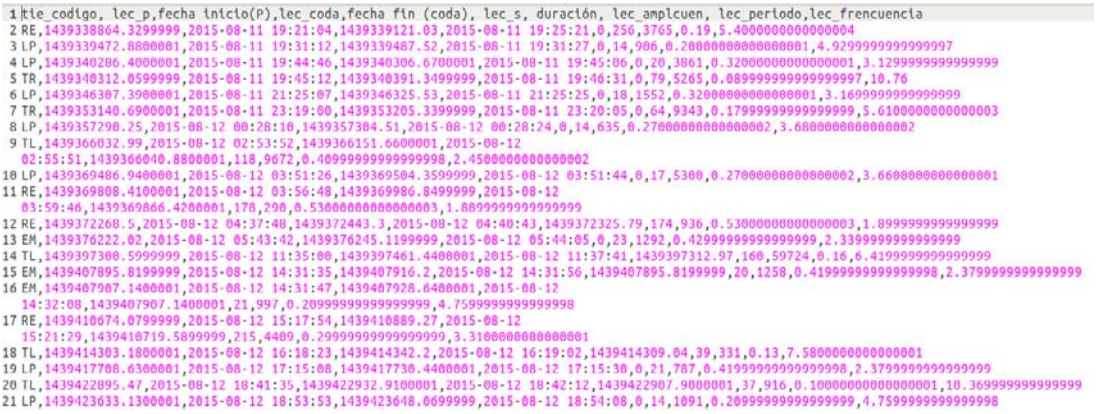


FIGURA 36 - Ejemplo de .csv

Tal y como se puede observar, en un archivo .csv no se incluye la información del día indicado, si no también parte del análisis del día anterior. Este hecho se detectó

por inspección ocular de los ficheros ya que en la documentación disponible no se especificaba.

Un análisis diario sobre un archivo .txt se compone de los 10 atributos que se detallan en la Tabla 4:

Atributo	Descripción	Formato
tie_codigo	este campo recoge el tipo de onda sísmica registrada, una de las 18 posibles registradas en torno al volcán Puracé	Nominal
lec_p	indica el instante en que llegó la primera onda P	Numérico
fecha de inicio (P)	indica cuándo se comenzó a registrar la onda sísmica	Numérico
lec_coda	indica la duración de la coda	Numérico
fecha fin (coda)	indica cuándo se dejó de registrar la onda sísmica	Numérico
lec_s	indica la duración del sismo, desde la onda P hasta el fin de la coda	Numérico
duración	expresa en segundos la duración de la onda sísmica	Numérico
lec_amplcuen	valor de la amplitud de la onda	Numérico
lec_periodo	período de la onda	Numérico
lec_frecuencia	frecuencia de la onda	Numérico

Tabla 4 - Atributos .csv

Como se puede observar en el fichero, hay dos fechas, fecha de inicio (P) y fecha fin (coda) por cada entrada, estando las ondas asignadas a intervalos de tiempo. En la primera instancia por ejemplo, se observa que durante el 11 de Agosto de 2015 se registró una onda sísmica RE entre las 19:21:04 y las 19:25:21.

En total, será posible registrar 18 tipos de ondas sísmicas diferentes en torno al volcán Puracé:

Tipo de onda sísmica	Código
Volcano Tectónico	VT
Híbrido	HB
Tornillo	TO
Tectónico Local	TL
Regional	RE
Distante	DS
Tremor	TR
Explosión en Mina	EM
Act. Superficial	SU
No Clasificable	NC
Calibración	CA
Rayo	RY
Explosión	EX

Avalancha	AV
Baja Frecuencia	BF
No Determinado	ND
Hielo	HI

Tabla 5 - Ondas sísmicas

4.3 Preparación de los datos

En la práctica sólo se realizaron modelos con los datos de agosto a diciembre del año 2015, los datos del mes de julio estaban incompletos al no tener asociados análisis .csv a los ficheros .txt con la monitorización diaria.

La carpeta del año 2016 no contenía datos, tan sólo los análisis posteriores y aunque se informó al Instituto de Geología y Minas no se solucionó.

Sin embargo, con los ficheros de datos facilitados por el Instituto no se pueden crear los modelos de clasificación directamente, pues los ficheros .txt con la información diaria no incluyen el tipo de onda sísmica.

Fue necesario elaborar un programa que relacionara cada fichero .txt con cada fichero .csv, se le añadió una cabecera especial y modificó su extensión .txt a .arff para ser procesado por WEKA. Este proceso se realizó fichero .txt a fichero .txt y se refleja en la Fig.38:

Registros diarios asociados a .csv (formato.txt)

Análisis diario .csv Registro diario con ondas .arff

FIGURA 37 - Diagrama de conversión de ficheros

El fichero .arff resultante contenía todas las instancias del .txt la onda indicada por el .csv y la cabecera del fichero .arff correspondiente.

4.3.1 Elaboración del fichero maestro

Ninguno de los ficheros .arff generados contenía los 18 tipos de ondas sísmicas que se podían dar. Por esta razón, fue necesario crear un fichero maestro que contuviera todos los tipos de onda registrables.

Este fichero se elaboró por inspección ocular examinando los ficheros .csv diarios desde Agosto de 2015 hasta Diciembre de 2015. En total se detectaron 11 de las 18 ondas mencionadas y la ausencia de mediciones del sismógrafo de Pilimbala, por lo que el fichero maestro no contiene referencia alguna a este sismógrafo ni en sus instancias ni en la cabecera de Weka y tiene tan sólo 15 atributos, correspondientes a las mediciones del resto de sismógrafos.

Atributo		**Formato**
Fecha (día y hora)		Fecha
Agua Blanca (ABL)	E	Numérico
	N	Numérico
	Z	Numérico
Coucy (COC)	E	Numérico
	N	Numérico
	Z	Numérico
Condor (CON)	E	Numérico
	N	Numérico
	Z	Numérico
Lavas Rojas (LAR)	E	Numérico
	N	Numérico
	Z	Numérico
Shaka (SHA)	E	Numérico
	N	Numérico
	Z	Numérico

Tabla 6 - Atributos del Conjunto Maestro

Para elaborar el fichero maestro se juntó la información de los días 221, 235, 244 y 248 de todos los días disponibles. El día 221 contenía 9 de los 11 tipos de onda detectados y los ficheros 235, 244 y 248 instancias de al menos uno de los dos tipos de onda ausentes. El resto de los días inspeccionados registraban una actividad sísmica muy similar entre sí.

Se consideró apropiado eliminar el atributo date por considerar que la información que pudiera aportar este atributo se encontraba contenida en el resto de atributos.

En caso de que faltara el valor de un atributo se consideró oportuno que Weka infiriera su valor en tiempo de ejecución, dado que el tamaño final del conjunto era de 32,9MB. Este tamaño se consideró adecuado para poder ser manejado por otras herramientas en caso de necesidad, como por ejemplo Excel y para poder incluir ondas de los que no era posible obtener más registros. Con el propósito de poder

aplicar la mayoría de algoritmos posibles se asignó a cada tipo de onda un valor en código binario. En la Tabla 6 se puede ver la asignación:

Tipo de Onda	Código Binario
VT	00000
LP	00001
HB	00010
HI	00011
TO	00100
TL	00101
RE	00110
DS	00111
TR	01000
EM	01001
SU	01010
NC	01011
CA	01100
RY	01101
EX	01110
AV	01111
BF	10000
ND	10001
NULL	11111

Tabla 7 - Correspondencia entre Ondas y su Código Binario

Como se puede observar se ha añadido un tipo de onda NULL a las 18 originales. Esta asignación fue necesaria porque en la monitorización diaria, los sismógrafos registraban ondas mecánicas que no se correspondían con ninguno de los tipos de onda indicados. Se estimó relevante, identificar este tipo de ondas y agruparlas en una misma clase para evitar que fueran interpretadas como ruido.

Las ondas que pasaron a formar parte del fichero maestro fueron finalmente 10 de las originales y la onda NULL:

Tipo de Onda	Código Binario
VT	00000
LP	00001
TO	00100
TL	00101
RE	00110
DS	00111
TR	01000

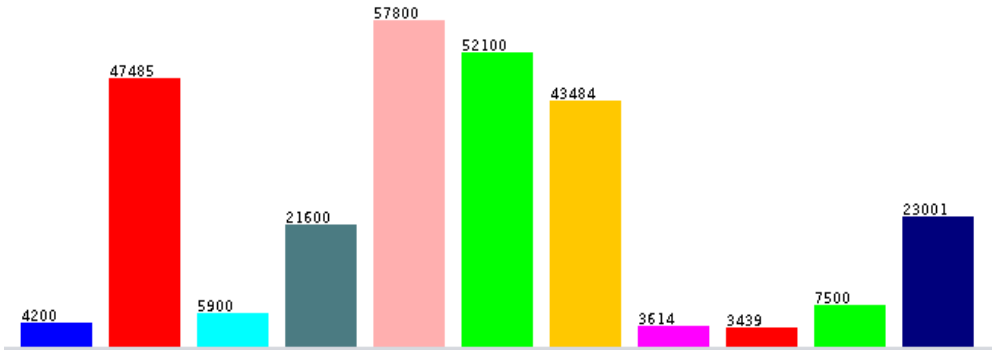
SU	01010
NC	01011
ND	10001
NULL	11111

Tabla 8 - Ondas incluidas en el fichero maestro

4.3.2 Fichero maestro: artificial_data.arff

El fichero maestro se elaboró usando la información contenida de los días 221, 235, 234 y 248 alcanzando un tamaño final de 32,9MB.

Su nombre final fue artificial_data.arff y contenía instancias asociadas a las ondas anteriores (véase Tabla7).



Onda	Número de Instancias
VT	4200
LP	47485
TO	5900
TL	21600
RE	57800
DS	52100
TR	43484
SU	3614
NC	23439
ND	7500
NULL	23001

FIGURA 38 – Número de isntancias en el fichero maestro

Como se puede observar, las clases tienen un número de instancias muy distinto entre sí. Algunas de las ondas cuya clasificación se perseguía aparecían en uno o dos ficheros por lo que no era posible añadirlas empleando otros ni se estimó apropiado prescindir de ellas por los que se optó por dejar que Weka balanceara las clases. Tras

realizar estas operaciones se creó un nuevo fichero y se estableció como conjunto de entrenamiento para elaborar clasificadores.

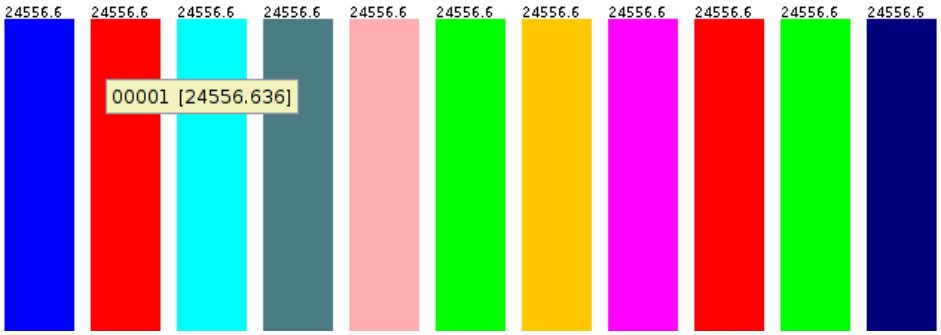


FIGURA 39 – Número de instancias en los conjuntos Conjunto 1 y Conjunto 2

4.3.3 Elaboración de subconjuntos

Sobre el fichero maestro se realizaron análisis exploratorios sobre los datos del fichero maestro para detectar los atributos más relevantes ejecutando las siguientes técnicas:

- **Evaluación por subconjuntos de clasificadores:** evalúa la valía de un subconjunto de atributos en base a su habilidad predictiva sobre la clase y el nivel de redundancia entre ellos.
- **Análisis de correlación entre atributos:** evalúa la valía de un atributo en baso a la correlación de Pearson entre la clase y él.
- **Evaluación del rato de ganancia de un atributo de cara a la clasificación:** evalúa la valía de un atributo midiendo su ratio de ganancia respecto a la clase.
- **Información del ratio de ganancia de un atributo respecto a la clasificación:** evalúa la valía de un atributo en base a la información aportada para llevar a cabo la clasificación.

La evaluación por subconjuntos especifica si un atributo es o no relevante con una probabilidad de 0% a 100% mientras que el resto de clasificadores elaboran un *ranking*, en el que otorgan a cada atributo un valor numérico. Al diferir los resultados de las técnicas, se elaboraron dos subconjuntos de datos, uno empleando los atributos detectados como relevantes por la evaluación por subconjuntos de clasificadores y otro, agrupando los atributos comunes en el resto de técnicas siempre y cuándo estuvieran en la mitad superior del ranking de cada uno de ellos:

Evaluación por subconjuntos de clasificadores	Atributos relevantes y comunes detectados en los anteriores análisis
SHA_N	**X**
ABL_E	**X**
COC_N	**X**
CON_E	**X**
CON_N	CON_N

LAR_E	LAR_E
LAR_N	LAR_N
SHA_E	SHA_E
SHA_Z	SHA_Z
X	CON_Z
X	COC_Z

Tabla 9 – Atributos de los conjuntos Conjunto 1 y Conjunto 2

Como se puede observar en la Tabla 9, el número de atributos es diferente en uno y otro subconjunto. En el subconjunto elaborado a partir de la evaluación por subconjuntos de clasificadores, hay 9 atributos y en el otro 7 de los cuales 5 son comunes con el otro.

En total, se realizarán modelos con el conjunto maestro (artificial_data.arff) y estos dos subconjuntos que en adelante se llamarán Conjunto de datos 1 y Conjunto de datos 2 para examinar con cuál de ellos se obtienen los mejores resultados.

4.4 Obtención de Modelos

En esta fase se realizaron experimentos de clasificación y posteriormente de segmentación para contrastar los resultados de los clasificadores.

Las técnicas que se emplearon en la clasificación fueron C4.5, Hoeffding Tree, Naïve Bayes y Red Bayesiana.

4.4.1 Algoritmos de clasificación

La clasificación o inducción supervisada tiene como objetivo analizar un conjunto de datos y generar un modelo que permita distinguir las clases predefinidas contenidas en él para clasificar correctamente instancias futuras que no dispongan del atributo de clase.

4.4.1.1 Árboles de decisión

A partir del conjunto de datos de entrenamiento, un árbol de decisión elabora un modelo que permite en base al valor de los atributos clasificar cualquier instancia de ese conjunto (y futuras) en su clase o categoría correspondiente. Es una técnica de aprendizaje supervisado.

Su objetivo es separar con la mayor exactitud las clases representadas en el conjunto de datos. Esta tarea se lleva a cabo a través de preguntas que se formulan de forma ordenada, en la que la respuesta determina a la pregunta actual determinará la siguiente. Estas preguntas se pueden responder como si/no, verdadero/falso o valor (propiedad) en un conjunto de valores. (Blázquez García 2004).

Un árbol se compone de un nodo padre o raíz que por convenio se muestra en la parte superior. Este nodo se ramifica en varios nodos hijos de acuerdo al factor de ramificación que se use y que es constante en todo el árbol. Los nodos hijos pueden a su vez ramificarse o no, en caso de no ramificarse son al mismo tiempo un nodo hijo y

un nodo hoja. Un nodo hijo evalúa un determinado atributo y determina en base al mismo el siguiente nodo por el que continuar el proceso de clasificación. Un nodo hoja es un nodo final, que no se ramifica en ningún otro y que recoge uno de los posibles valores del atributo de clase.

Los nodos representan el nombre del atributo e incluyen la pregunta que se ha de contestar sobre ese atributo para continuar con la clasificación de estar tratando con un nodo hijo, los arcos representan la respuesta recibida para esa pregunta y conectan con el siguiente nodo y la siguiente pregunta, los nodos hojas representan las clases.

Navegar por el árbol y llegar a un determinado nodo hoja implica que la instancia que se quiere clasificar ha cumplido con todos los requisitos de la clase que señala el nodo hoja.

Los dos árboles empleados en esta fase son C4.5 y Hoeffding Tree.

C4.5

El árbol ID3 (*Induction of Decision Tree 3*) presentado por Quinnlan en 1979 basado en el trabajo de Hunt (1962) fue mejorado con heurísticas que resultaron en el algoritmo C4.5 (1993). Sobre esta versión se aplicaron otros cambios que resultaron en el algoritmo C5.0. En la herramienta empleada para utilizar este trabajo, Weka, C4.5 recibe el nombre de C4.5. (Hssina B. s.f.)

El funcionamiento de ID3 es simple. A partir de un conjunto de datos de entrada al que considera una descripción exacta de la realidad crea un clasificador empleando un árbol de decisión. En su origen, se pensó para ser utilizado por datos nominales aunque en la actualidad es posible tratarlo para datos con valores numérico si se pre-procesan agrupándolos en intervalos etiquetados de forma nominal. Su factor de ramificación recibe el nombre de B_j y se calcula variable por variable en base al número de atributos discretos que contiene de la variable *j*. La profundidad del árbol coincide con el número de atributos del conjunto de datos. (Quinnlan 1985). C4.5 puede tratarse sin pre-procesamiento sobre conjuntos de datos numéricos, nominales o mixtos. Incluyen heurísticos que aportan mayor información sobre el conjunto de datos y permiten elaborar un algoritmo de construcción recursivo. Estos heurísticos son el atributo informativo **(**los nodos se ordenan en baso a la información que ofrecen sobre el conjunto de datos de mayor a menor, siendo el atributo más informativo el nodo padre) y la entropía (la información que ofrece cada nodo se mide en base al valor de su entropía, es decir, en base a la ganancia de información tras evaluar el atributo. (Fernández Rebollo 2015)):

- Atributo informativo: los nodos se ordenan en baso a la información que ofrecen sobre el conjunto de datos de mayor a menor, siendo el atributo más informativo el nodo padre.

- Entropía: la información que ofrece cada nodo se mide en base al valor de su entropía, es decir, en base a la ganancia de información tras evaluar el atributo. (Fernández Rebollo, 2015)

Este algoritmo recibe como parámetros de entrada el atributo categórico *C*, que indica la clase; un conjunto de atributos no categóricos *R*, que son los demás atributos de la tabla de los datos a tratar; y un conjunto de entrenamiento *S*, que contiene los ejemplos de la tabla. C4.5 verifica cuál es el atributo *D* más informativo de *R*, o sea, con mayor ganancia de información para el conjunto de entrenamiento *S*. Entonces, subdivide este conjunto S, de acuerdo con cada uno de los valores de este atributo más informativo *D* (Feldens, 1997).

Este proceso se repite de forma recursiva hasta que se cumple alguna de las siguientes tres condiciones de parada:

a) El conjunto de ejemplos ha sido tratado en toda su extensión.
b) Todos los ejemplos del conjunto de entrenamiento S son de la misma clase de C.
c) No quedan más atributos no categóricos de R para la clasificación.

Hoeffding Tree

Este árbol se basa en el árbol ID3 desarrollado por Quinland en 1973. Hoeffding tree (VFDT) es un algoritmo incremental que puede emplearse en sistemas off-line y online capaz de manejar grandes flujos de datos siempre y cuando éstos se generen respetando un determinado patrón en el tiempo.

Debe su nombre a la desigualdad de Hoeffding que calcula el número de observaciones (ejemplos, instancias, filas) necesarios para obtener estadísticos de forma precisa, en este caso, la bondad de un atributo (capacidad del atributo de influir en la clasificación) para colocarlo en una determinada posición en el árbol. Estos árboles son altamente eficientes y son capaces de inferior a partir de una pequeña muestra qué atributos conviene emplear para elaborar el modelo de clasificación.

Otra característica muy interesante de este algoritmo es, que la salida, el modelo generado es prácticamente idéntico asintóticamente al de otros árboles de decisión no incrementales siempre y cuando éstos dispongan de una gran batería de ejemplos (Kirby s.f.).

4.4.1.2 Métodos de inferencia bayesianos

A la hora de realizar predicciones, la probabilidad era la herramienta que matemáticos y estadísticos empleaban para tratar la incertidumbre inherente a ellas. El filósofo inglés Thomas Bayes (1702-1761) elaboró **el Teorema de Bayes** que describe la probabilidad de que un evento tenga lugar, basándose en las condiciones en las que se dio ese evento en el pasado.

El teorema de Bayes tiene múltiples aplicaciones siendo una de ellas la inferencia estadística empleada para averiguar la probabilidad de que un determinado evento tenga lugar. El razonamiento bayesiano proporciona un enfoque probabilístico a la inferencia (Blázquez García 2004). En la Fig.41 el teorema de Bayes:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

$P(A)$ es la probabilidad a priori de A, es decir, es la medida de verosimilitud que se tiene de que el suceso $P(A)$ocurre.

- $P(B)$ se define análogamente a $P(A)$.
- $P(B|A)$ es la probabilidad de observar el suceso B si ocurre A.
- $P(A|B)$ es la probabilidad a posteriori de que ocurra A si ha ocurrido B.

FIGURA 40 - Teorema de Bayes

En los problemas en los que se practica la inferencia bayesiana A es una hipótesis y B son los datos de muestra. Ante un conjunto de posibles hipótesis, se escogerá aquella cuya probabilidad $P(A|B)$ o análogamente $P(hipótesis|DATOS)$ sea mayor. Esta hipótesis recibe el nombre de MAP (*máxima hipótesis a posteriori*).

Redes bayesianas

Las redes bayesianas son grafos acíclicos en los que cada nodo representa una variable y un arco una dependencia probabilística en la que se especifica la probabilidad condicionada de cada variable respecto a sus padres. El sentido del arco determina quién es la variable dependiente y cuál la independientemente siendo éstas la situada en el extremo del arco y en el origen respectivamente.

Las redes bayesianas son susceptibles de dos interpretaciones:

a) **Distribución de probabilidad:** en este primer caso, la red bayesiana representará la probabilidad conjunta de todo el conjunto de datos variable a variable.
b) **Base de reglas:** el centro de atención serán los arcos, siendo ellos los que cuantifiquen las probabilidades entre nodos (variables) conectados entre sí.

Las redes bayesianas permiten aprender sobre relaciones de dependencia y causalidad, permiten combinar conocimiento con datos, disminuir el sobreajuste en los modelos y manejar BBDD incompletas.

Las redes bayesianas se construyen combinando dos tipos de aprendizaje:

a) **Aprendizaje paramétrico:** tiene como objetivo obtener las probabilidades y probabilidades condicionales de cada variable.
b) **Aprendizaje estructural:** establece las relaciones de dependencia e independencia entre las variables del conjunto de estudio. Este aprendizaje depende de la estructura de la red (árbol, poliárbol, red multiconectada)
c) **Aprendizaje paramétrico o estructural y otro tipo de aprendizaje:** la estructura viene dada y los estadísticos de esta técnica se emplean para mejorar el modelo.

4.4.2 Experimentos de clasificación

En esta sección se describe la ejecución de los algoritmos C4.5, Hoeffding Tree, Naïve Bayes y Red Bayesiana aplicando la técnica de validación cruzada con 10

particiones en los tres conjuntos de datos: conjunto inicial o maestro, conjunto 1 y conjunto 2.

En los epígrafes de esta sección se presentarán los resultados de cada algoritmo sobre cada uno de los conjuntos y los valores de la curva PRC de todos los clasificadores.

La curva PRC (*Precision-Recall Curve*) es útil en tareas como la clasificación en el que es importante estimar si un valor es relevante o no. El término ***precision*** (en español precisión) es la fracción de instancias relevantes mientras que el término ***recall*** (en español memoria) hace referencia al número de instancias relevantes que se han recuperado de todo conjunto.

Entre dos algoritmos que emplean el mismo conjunto de datos si los valores PRC son altos, escogeremos el algoritmo con menor ***recall*** puesto que ha sido capaz de obtener buenos resultados con menor información que su competidor (Saito T 2015) (StackExchange 2014).

Conjunto Inicial (fichero maestro):

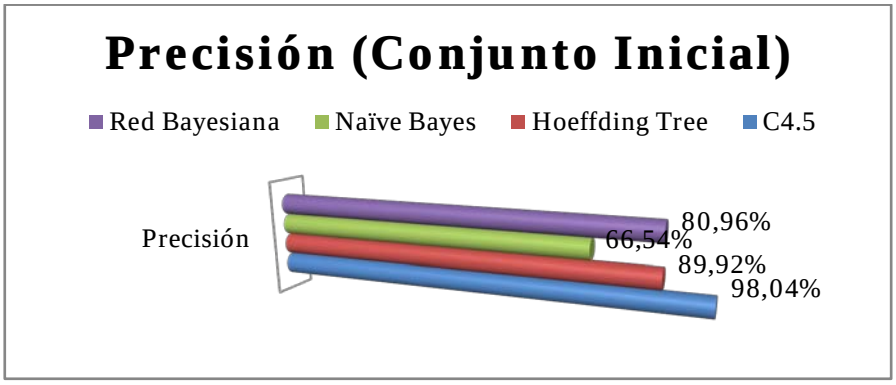


FIGURA 41 - Resultados de los Clasificadores con el Fichero Maestro

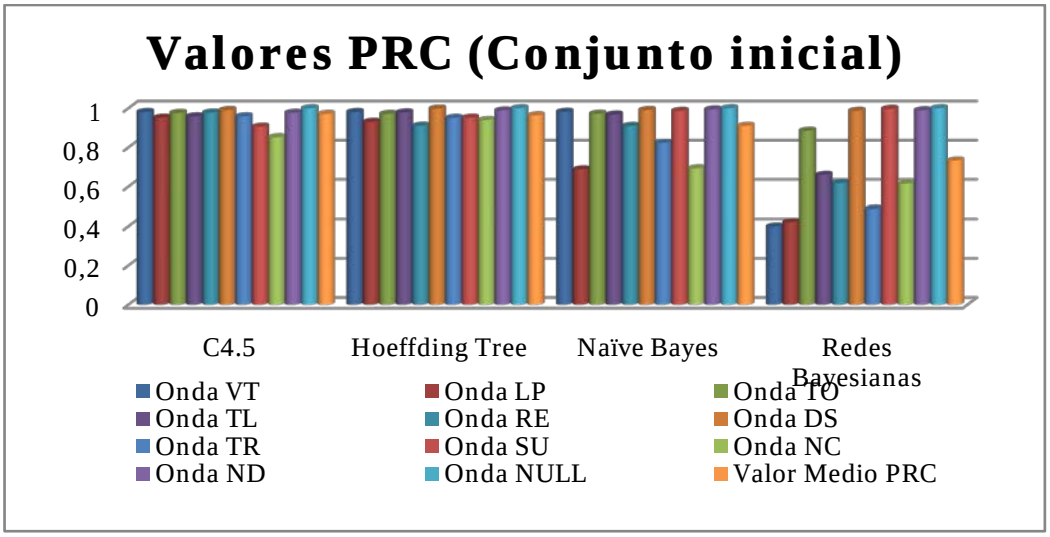


FIGURA 42 - Valores PRC (Conjunto inicial)

Conjunto 1:

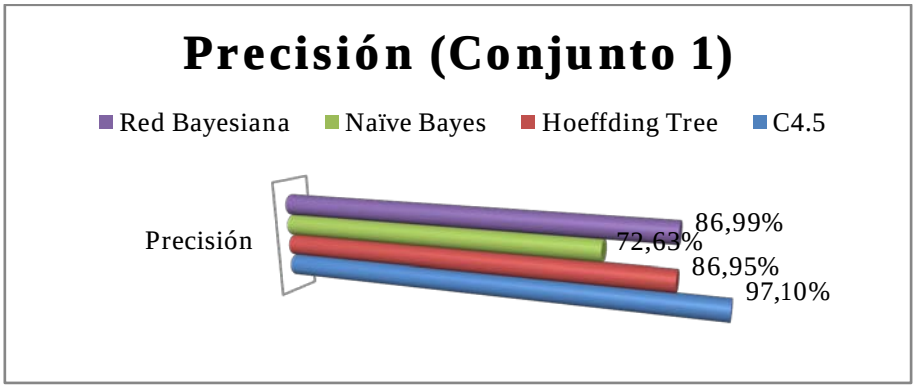


FIGURA 43 - Resultados de los Clasificadores con Conjunto 1

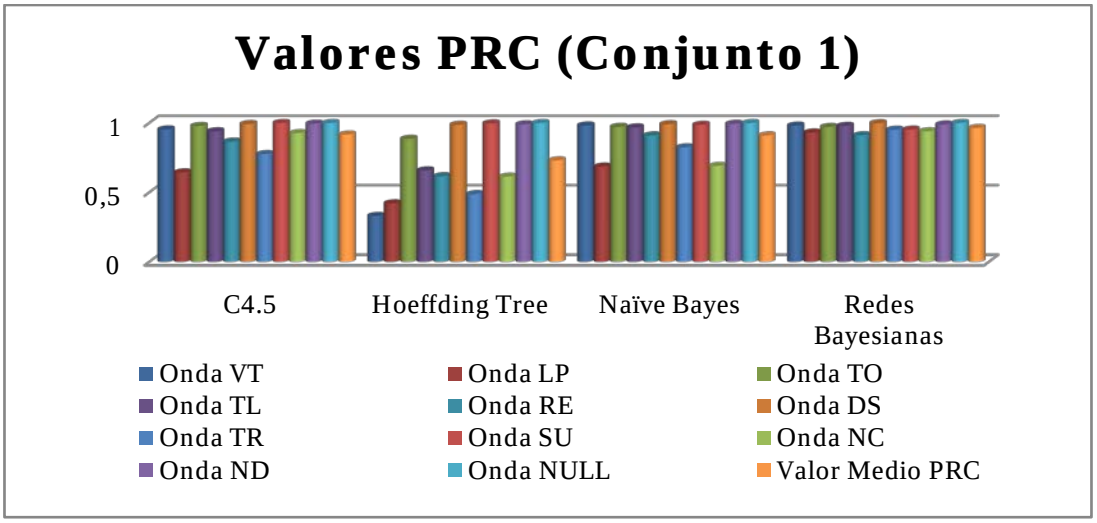


FIGURA 44 - Valores PRC (Conjunto 1)

Conjunto 2:

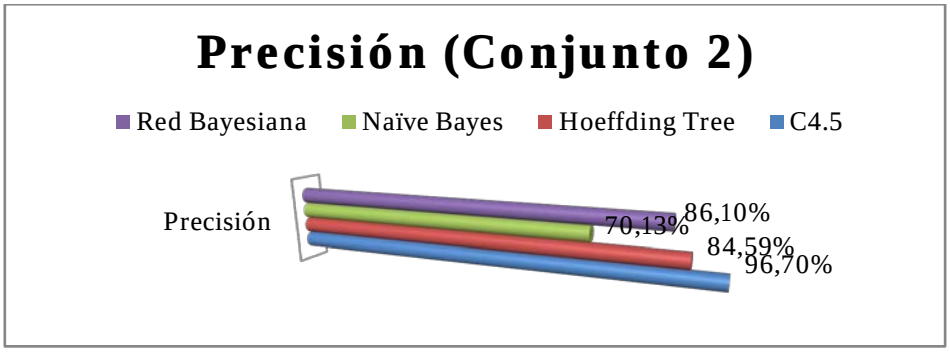


FIGURA 45 - Resultados de los Clasificadores con Conjunto 2

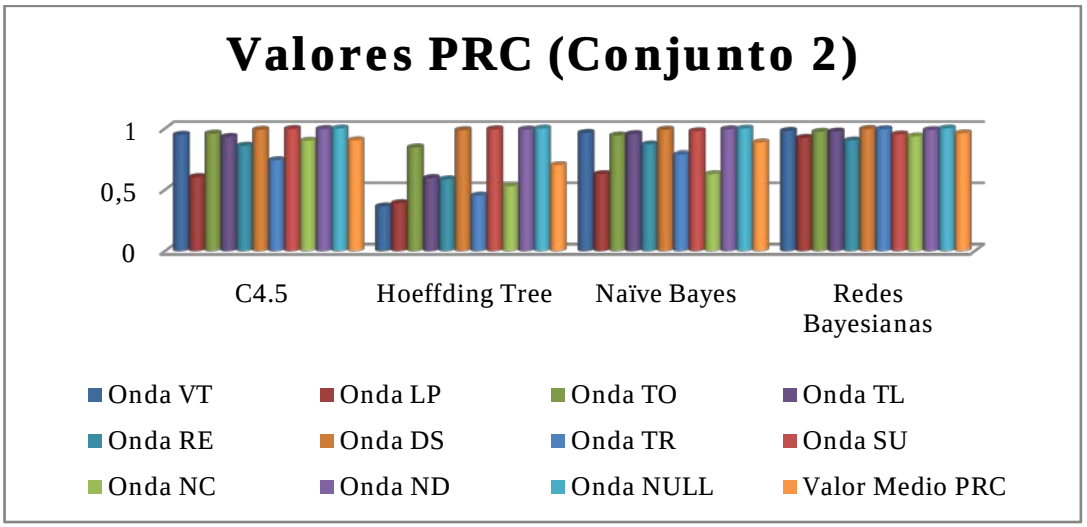


FIGURA 46 - Valores PRC (subconjunto 2)

4.4.3 Evaluación de los modelos de Clasificación

Como se puede observar, en todos los casos se siguen unas distribuciones muy parecidas como se aprecia en el siguiente gráfico y la siguiente tabla:

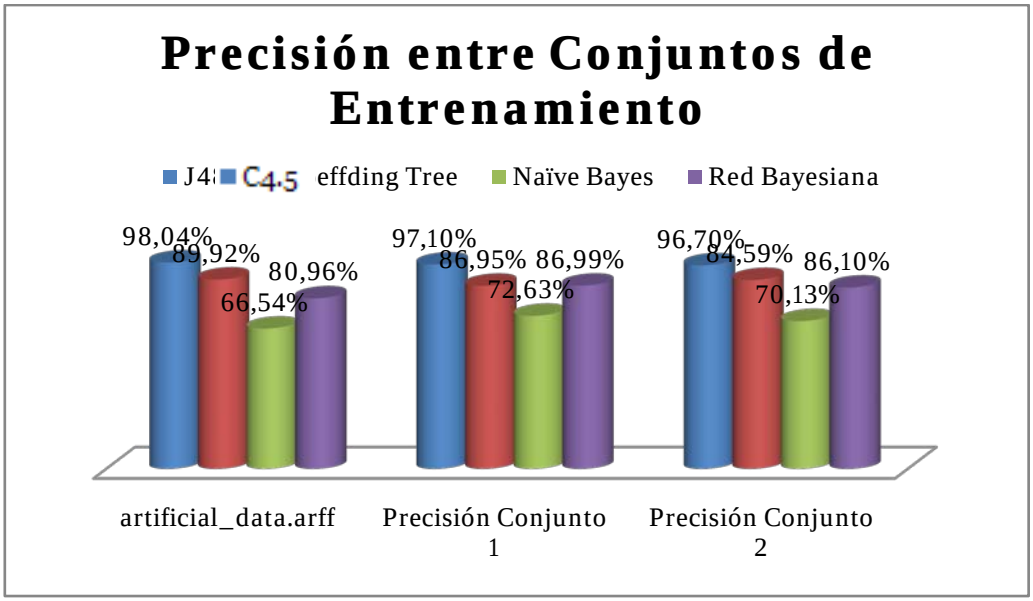


FIGURA 47 – Precisión entre Conjuntos de Entrenamiento

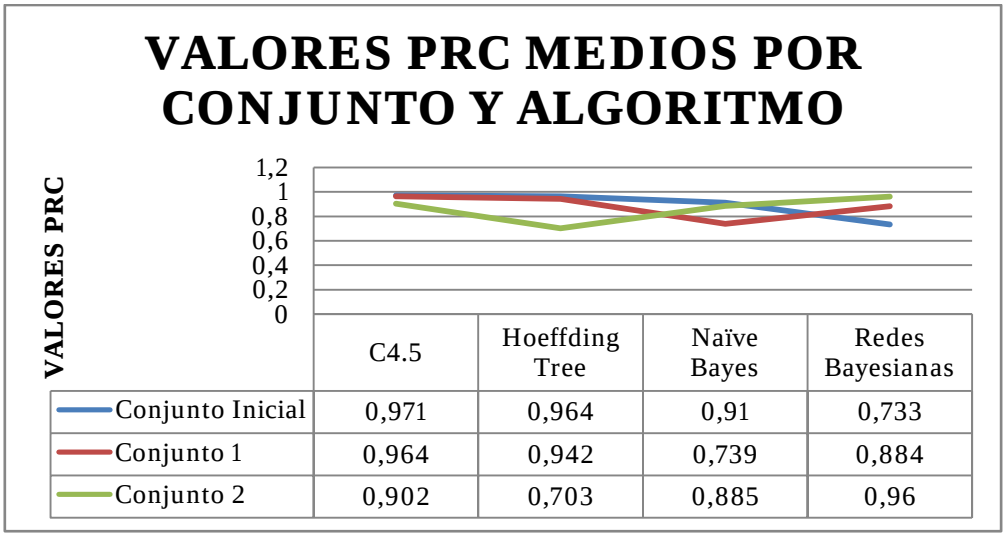
	C4.5	Hoeffding Tree	Naïve Bayes	Red Bayesiana
Precisión Conjunto Maestro	98,04 %	89,92%	66,54%	80,96%
Precisión Conjunto 1	97,10%	86,95%	72,63%	86,99%
Precisión Conjunto 2	96,70%	84,59%	70,13%	86,10%

Tabla 10 - Tabla de Valores

El mejor conjunto de datos parece ser, el Conjunto 2 ya que los clasificadores tienen valores altos muy parecidos entre sí. El mejor algoritmo, sería el árbol generado por C4.5 modelado a partir del conjunto de datos inicial.

Sin embargo, estos resultados pueden ser engañosos por el sesgo que hay en ellos al haber sido balanceados y además ¿el hecho de que un clasificador clasifique correctamente implica que está clasificando con la misma precisión para todos los valores del atributo clase o sólo para unos pocos?

Evaluar los valores PRC nos dará más información, véase la Tabla 10, a continuación:



	C4.5	Hoeffding Tree	Naïve Bayes	Redes Bayesianas
Conjunto Inicial	0,971	0,964	0,91	0,733
Conjunto 1	0,964	0,942	0,739	0,884
Conjunto 2	0,902	0,703	0,885	0,96

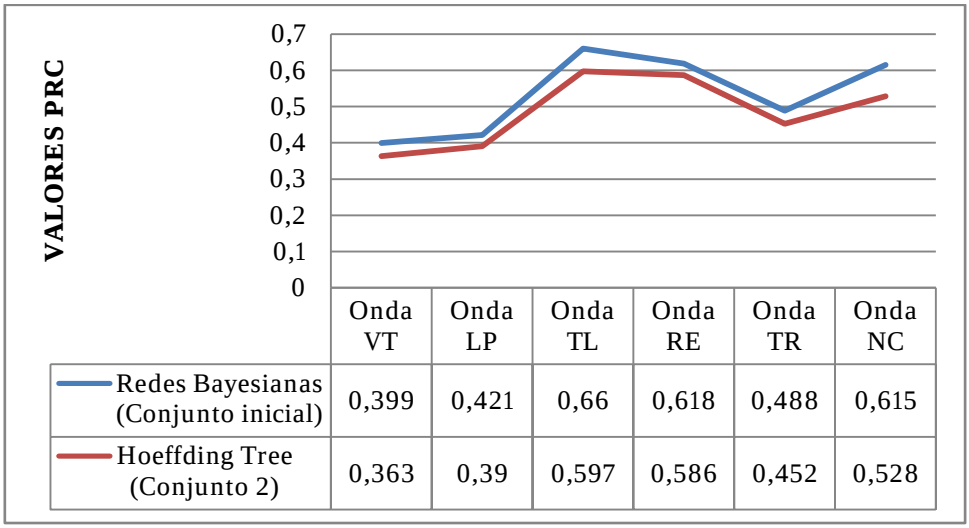
FIGURA 48 - Valores PRC medios por Conjunto y Algoritmo

El algoritmo con valores PRC más dispersos es Hoeffding Tree en los conjuntos 1 y 2. Esto indica que a pesar de sus altos valores de precisión no se aplican a las clases, VT, LP, TL, RE, TR y NC.

En el conjunto inicial, éste algoritmo es Redes Bayesianas y curiosamente para las mismas ondas que Hoeffding Tree para el conjunto 2.

De acuerdo, a la definición dada anteriormente de PRC, no sería preciso elaborar los conjuntos Conjunto 1 y Conjunto 2 ni balancearlos ya que con un pre-procesamiento mínimo (eliminar el atributo *date*) se obtienen buenos resultados. Desgraciadamente, no es posible adjuntar el árbol por su gran número de nodos.

En la siguiente tabla, se incluyen las ondas con los algoritmos cuyos valores PRC son bajos.



	Onda VT	Onda LP	Onda TL	Onda RE	Onda TR	Onda NC
Redes Bayesianas (Conjunto inicial)	0,399	0,421	0,66	0,618	0,488	0,615
Hoeffding Tree (Conjunto 2)	0,363	0,39	0,597	0,586	0,452	0,528

Tabla 11 – Valores PRC para Hoeffding Tree en el fichero maestro y Redes Bayesianas en el conjunto 2

Onda	Número de instancias en el Conjunto Inicial o Maestro	Número de instancias en el Conjunto 2
VT	4.200	24.556
LP	47.485	24.556
TL	21.600	24.556
RE	57.800	24.556
TR	43.484	24.556
NC	23.439	24.556

Tabla 12 - Ondas con valores PRC bajos y número de instancias por conjunto

Como se puede observar, hay un patrón muy claro en los valores PRC de ambos clasificadores para las ondas representadas, incrementándose su fiabilidad con el método de Redes Bayesianas.

Este hecho señala tres eventos: el primero, que el balanceo de clases es inútil. Prácticamente, la curva para las Redes Bayesianas y Hoeffding Tree es la misma, solo que la de éste último ha sido balanceada. Esto revela también que la naturaleza de los algoritmos que se ejecuten es importante.

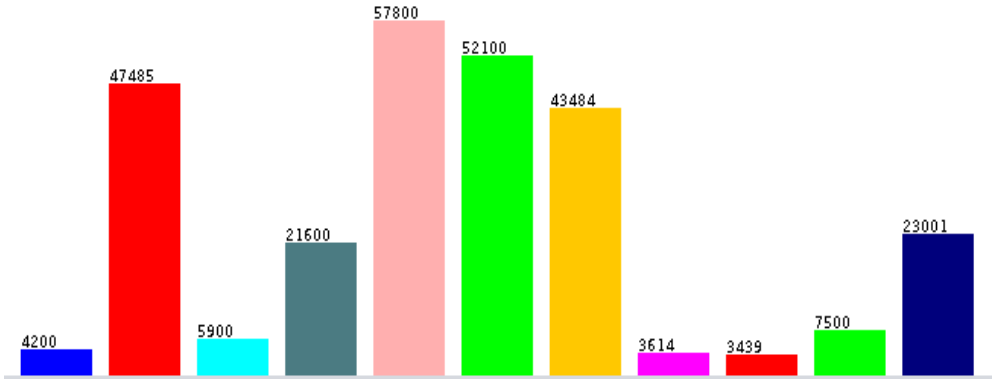
Las Redes Bayesianas al construirse empleando el Teorema de Bayes funcionan mejor a mayor número de ejemplos disponibles, en cambio, Hoeffding Tree empeora sus valores PRC puesto que, los ejemplos generados por Weka deben de ser muy similares entre sí y no aportan información.

Este hecho se subsanaría empleando un conjunto de datos que a priori contuviera el mismo número de ejemplos por cada valor de la clase o bien usando una herramienta que elaborara nuevas instancias con métodos que generaran instancias relevantes para el conjunto y las que eliminara no fueran relevantes.

4.4.4 Algoritmos de segmentación

Al margen de la detección de patrones sin restricciones ni conocimiento a priori sobre un conjunto de datos, las técnicas de segmentación resultan útiles para contrastar los resultados obtenidos en una tarea de clasificación.

Al haber obtenido tan buenos resultados en los clasificadores, resulta interesante emplear otras técnicas para comprobar su fiabilidad. Para ello, se llevarán a cabo experimentos de segmentación sobre el **conjunto de datos maestro** para extraer la mayor información posible de él y poder respaldar los resultados de los clasificadores. Recordamos la distribución inicial de los ejemplos por clase en el fichero:



Onda	Número de Instancias
VT	4200
LP	47485
TO	5900
TL	21600
RE	57800
DS	52100
TR	43484
SU	3614
NC	23439
ND	7500
NULL	23001

FIGURA 49 - Conjunto de entrenamiento para Segmentación

Los algoritmos que se emplearán son Canopy, LVQ (*Learnig Vector Quatization*) y SKMedias (*Simple K Medias*).

K-Medias

Esta técnica de segmentación emplea una serie de heurísticas basándose en el *k* número de clases presentes en el conjunto de datos o en el que hayamos decidido dividirlo. **Es una técnica de aprendizaje no supervisado,** sencilla y eficiente.

Este algoritmo funciona de la siguiente manera, tras elegir un valor de *k* se toman *k* valores aleatorios del conjunto de datos que se denominarán centros o centroides y en torno a los cuáles se agruparán el resto de instancias del conjunto.

Esta elección aleatoria de *k* se emplea en un principio para evitar una agrupación sesgada en los conjuntos. Una elección supervisada de *k* puede garantizar el recubrimiento de todo el conjunto de datos y una mayor convergencia pero también puede evitar que se obtenga conocimiento oculto en los datos, es por tanto que en función del objetivo convendrá adoptar uno u otro enfoque.

Tras esta selección, el resto de elementos del conjunto de entrenamiento se asocian a uno u a otro segmento basándose en la media de los valores del clúster y la distancia entre el objeto y el clúster.

Cada vez que se añada una nueva instancia a un clúster se recalculan las medias de los centros o centroides a las instancias presentes en los clústeres hasta que los valores se estabilizan.

Canopy

Este algoritmo divide los datos de entrada en regiones de proximidad (canopies) en forma de hiperesferas cuya extensión se determina a partir de un valor T1 que determina su región central. Una segunda distancia T2 (en la que se cumple, T2 < T1) se emplea para establecer cuántos canopies crea el algoritmo (Witten 2016).

Se emplea como método de inicialización para el algoritmo K-Medias o sobre grandes conjuntos de datos en los que aplicar otro tipo de algoritmo es impracticable.

Su ejecución se divide en dos fases en la que primero se crean los clústeres y después se asignan instancias:

- **Fase 1**: se crean los canopies y se le asigna un ejemplo aleatorio del conjunto como centro a cada uno de ellos. Una vez hecho, se calcula la distancia que separa al resto de instancias de los canopies creados. Si una instancia tiene una distancia mayor que T2 para todos los canopies creados se formará un nuevo canopy o se descartará.
- **Fase 2:** las instancias son asignadas a los canopies empleando el valor T1 de cada canopy y la distancia calculada entre la instancia y ese canopy. Es posible que una misma instancia pertenezca a dos canopies.

LVQ (Redes de cuantización vectorial)

Es la versión supervisada del método de los mapas de Kohonen (SOM). A partir del número de clases del conjunto se elaboran segmentos y se clasifican instancias de acuerdo a su vecindad. Inicialmente, los prototipos se distribuyen de forma aleatoria en el espacio de entrada (Valls 2017).

En ellas, **no existe el concepto de vecindario** al ser una técnica de aprendizaje supervisado, tan solo se modifica la célula ganadora. En SOM se modifica todo el vecindario, mientras que en el LVQ únicamente la célula ganadora. **La dirección de la célula puede variar**, acercándose o alejándose de una clase, en SOM siempre se acerca acercándose a los ejemplos de las clases que representan y alejándose de los de las clases contrarias. **La tasa de aprendizaje se decrementa con el tiempo**.

La primera acción que lleva a cabo este algoritmo es distribuir los prototipos aleatoriamente en el espacio de entrada. El número de prototipos que contenga una clase se puede definir como un único número para todos ellos o como una proporción respecto al número de elementos que se agrupen en ella.

Según se van introduciendo ejemplos, los prototipos se desplazan empleando la misma regla que en SOM:

$$\frac{d_{\mu_{ij}}}{d_t} = \alpha(t)\tau_j(t)(e_i(t) - \mu_{ij}(t))$$

En el que el valor τ_j varía de la siguiente manera:

$$\tau_j = \begin{cases} 1, si\ C_j\ es\ ganadora\ y\ pertenece\ a\ la\ misma\ clase\ que\ e_i \\ -1, si\ C_j\ es\ ganadora\ y\ pertenece\ a\ distinta\ clase\ que\ e_i \\ 0, si\ C_j\ no\ es\ ganadora \end{cases}$$

4.4.5 Experimentos de Segmentación

En esta sección, se realizaron segmentaciones con el algoritmo Canopy con el objetivo de agrupar las clases en 3 y 5 Clústeres para observar qué ondas guardaban mayor parecido entre sí, un tercio y aproximadamente la mitad de clases que el conjunto inicial.

Después, con 11 Clústeres empleando los algoritmos LVQ y SKMedias guardando el atributo de clase para comprobar si los resultados se aproximaban a los obtenidos por los clasificadores o por el contrario, apenas guardaban semejanza, lo cual podría indicar que el sesgo aplicado al balancear las clases era demasiado grande y no había suficientes instancias para realizar un análisis correcto de los datos.

Resultados de la Experimentación con Canopy con 3 y 5 Clústeres

Canopy es un algoritmo de aprendizaje no supervisado útil para extraer conocimientos de dominios poco conocidos. En las Figuras 50 y 51 se detallan los resultados del experimento llevado a cabo sobre el conjunto de datos empleando 3 clústeres:

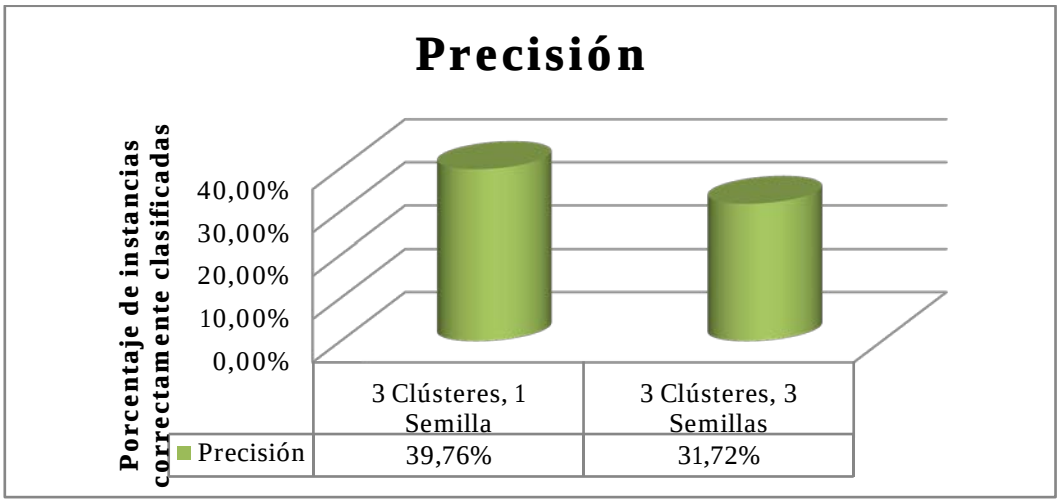


FIGURA 50 - Segmentación con 3 Clústeres (Canopy)

Las ondas en torno a las cuáles se organizan los clústeres son:

	Clúster 0	Clúster 1	Clúster 2

Canopy con 3 Clústeres, 1 semilla	RE (20%)	DS (19%)	TR (61%)
Canopy con 3 Clústeres, 3 semillas	VT (3%)	DS (58%)	TR (39%)

Tabla 13 - Correspondencia entre Clúster y Onda para Canopy con 3 Clústeres

Después de realizar estos experimentos, se procedió a repetirlos empleando 5 Clústeres:

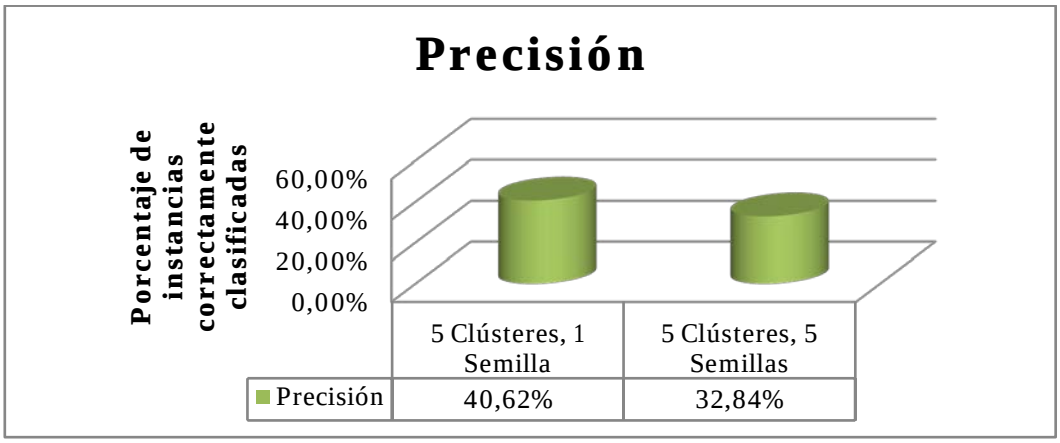


FIGURA 51 - Segmentación con 5 Clústeres (Canopy)

	Clúster 0	Clúster 1	Clúster 2	Clúster 3	Clúster 4
Canopy con 5 Clústeres, 1 semilla	LP (1%)	RE (19%)	DS (20%)	TR (60%)	SU (0%) ≈ 1133 *instancias*
Canopy con 5 Clústeres, 5 semillas	LP (4%)	RE (2%)	DS (55%)	TR (36%)	TO (2%)

Tabla 14 - Correspondencia entre Clúster y Onda para Canopy con 5 Clústeres

En ambos casos, se realizan clústeres para las ondas RE, DS y TR lo que deja claro que entre estos tres tipos de onda hay bastante diferencia y que el resto de ondas restantes se aproximarán a uno u a otro grupo. Como se puede observar, a mayor número de semillas, mayor es el tanto por ciento de instancias que se agrupan en torno a DS y TR y algo menor en torno a RE.

Resultados de la experimentación para LVQ y SKMedias para 11 Clústeres

El primer algoritmo en ser ejecutado fue el algoritmo LVQ seguido posteriormente del algoritmo SKMedias variando en el primer caso, la razón y en el segundo, el número de iteraciones y el heurístico (el método de elaboración de clústeres).

LVQ

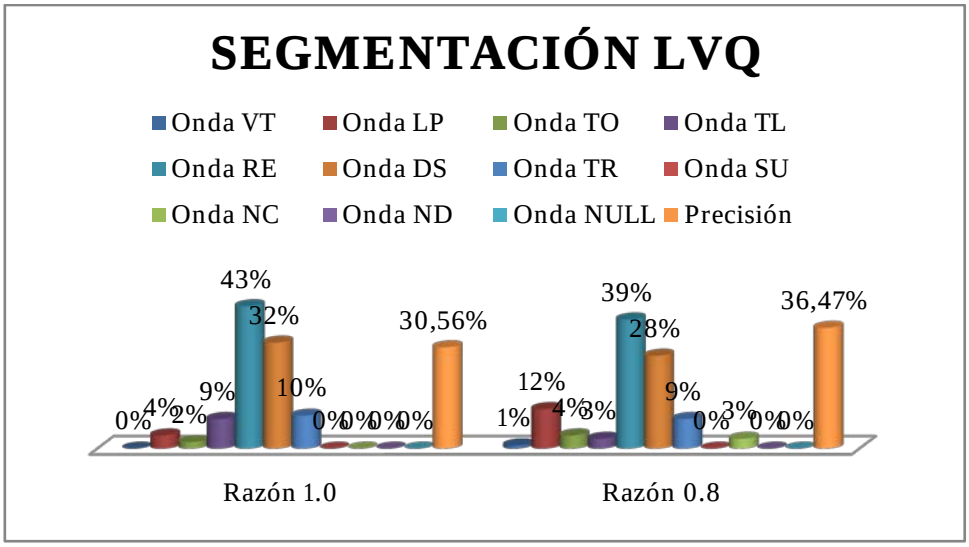


FIGURA 52 - Segmentación LVQ

En el primer experimento, las ondas que se quedan sin clúster asociados son: VT, SU, NC, ND y NULL. En el segundo, SU, ND y NULL.

Llama la atención, que las ondas con los valores más altos corresponden a las ondas detectadas por Canopy: RE, DS y TR cuando la razón es 1.0 y RE, DS, SU cuándo es 0.8.

	LVQ 1.0	**LVQ 0.8**
Onda RE	43%	39%
Onda DS	32%	28%
Onda TR	10%	X
Onda SU	X	12%

Tabla 15 - Clústeres con Mayor Densidad para LVQ

Otro hecho llamativo es, que se generan clústeres para todas las ondas pero no se asignan a todas ellas. En el primer caso, para las ondas VT, NC, ND, NULL y en el segundo para las ondas ND y NULL.

	LVQ 1.0	**LVQ 0.8**
Clúster 0	0% ≈ 41	0% ≈ 262
Clúster 1	0% ≈ 5	0% ≈ 22
Clúster 2	0% ≈ 356	X
Clúster 3	0% ≈ 160	X
	Onda VT, ND, NC, NULL	Onda ND, NULL

Tabla 16 - Clústeres sin asignar y Ondas sin segmentar para LVQ

Observando las matrices de confusión se observa como la densidad en torno a las ondas RE, DS, TL, TR es mayor que en el resto de ondas para LVQ con 1.0 y para LVQ con 0.8 para LP, RE y DS.

SKMEDIAS

Los experimentos realizados con SKMedias se realizarán en dos fases, primero aquellos cuya heurística empleada sea la distancia euclídea y posteriormente aquellos en los que sea la distancia Manhattan. Los resultados se presentan en el siguiente gráfico:

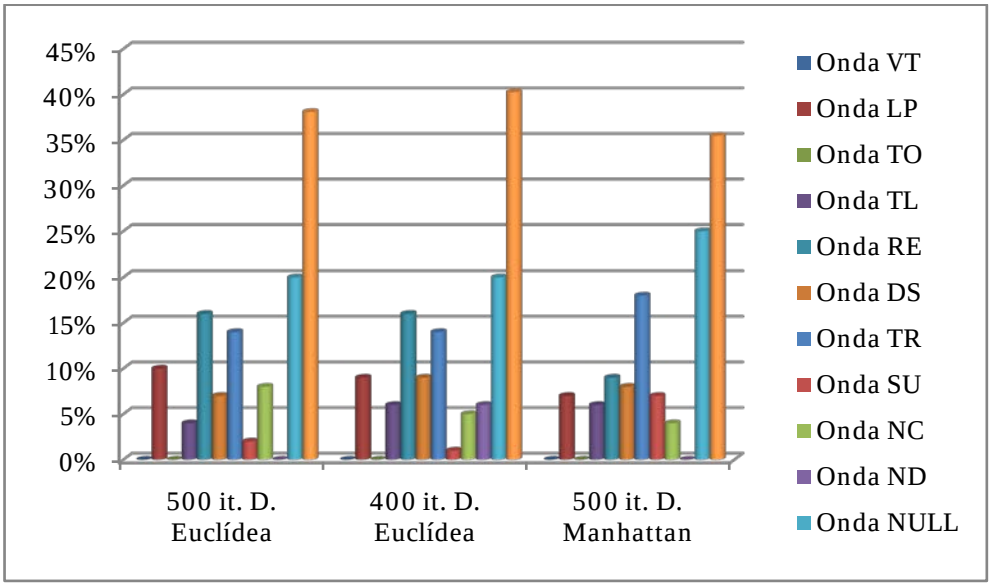


FIGURA 53 - Resultados para SKMedias

En este gráfico sólo se muestran los resultados más relevantes de la experimentación. SKMedias se ejecutó con el heurístico de distancia Euclídea para 300 y 200 iteraciones sin que los datos sufrieran variación alguna, al igual que para 400 iteraciones y el heurístico de distancia Manhattan, por lo que no se han incluido en la imagen.

Los gráficos muestran que las ondas NULL, TR, RE y LP reúnen ciertas características que las diferencia del resto y hace posible su agrupación siendo estos resultados consistentes hasta ahora con los experimentos realizados.

	500 it. Distancia Euclídea	**400it. Distancia Euclídea**	**500 it. Distancia Manhattan**
Onda LP	10%	9%	7%
Onda RE	16%	16%	9%
Onda TR	14%	14%	18%
Onda NULL	20%	20%	25%

Tabla 17 - Clústeres densos y Ondas Asociadas

Las ondas cuya precisión es de un 0% son ondas que carecen de clúster pero, el hecho de no haber podido asociarles un clúster no ha impedido su formación. Es decir, SKMedias ha agrupado instancias que guardan cierta semejanza entre sí pero a las que no ha podido asignar clúster algunos. La proporción de instancias respecto al total

asignadas a cada clúster se recogen a continuación así como las posibles ondas a las que podrían pertenecer los clústeres:

	Clúster 0	Clúster 1	Clúster 2
500 it. D. Euclídea	5%	7%	5%
400 it. D. Euclídea	7%	5%	X
500 it. D. Manhattan	6%	4%	7%
	Ondas VT, TO, ND	Ondas VT, TO, ND	Ondas VT Y TO

Tabla 18 - Clústeres sin asignar y Ondas sin segmentar

Comparación de resultados

Terminada ya la experimentación, se elaboró un gráfico en el que se recogían todos los experimentos realizados (incluidos aquellos etiquetados como no relevantes). Véase Fig. 54:

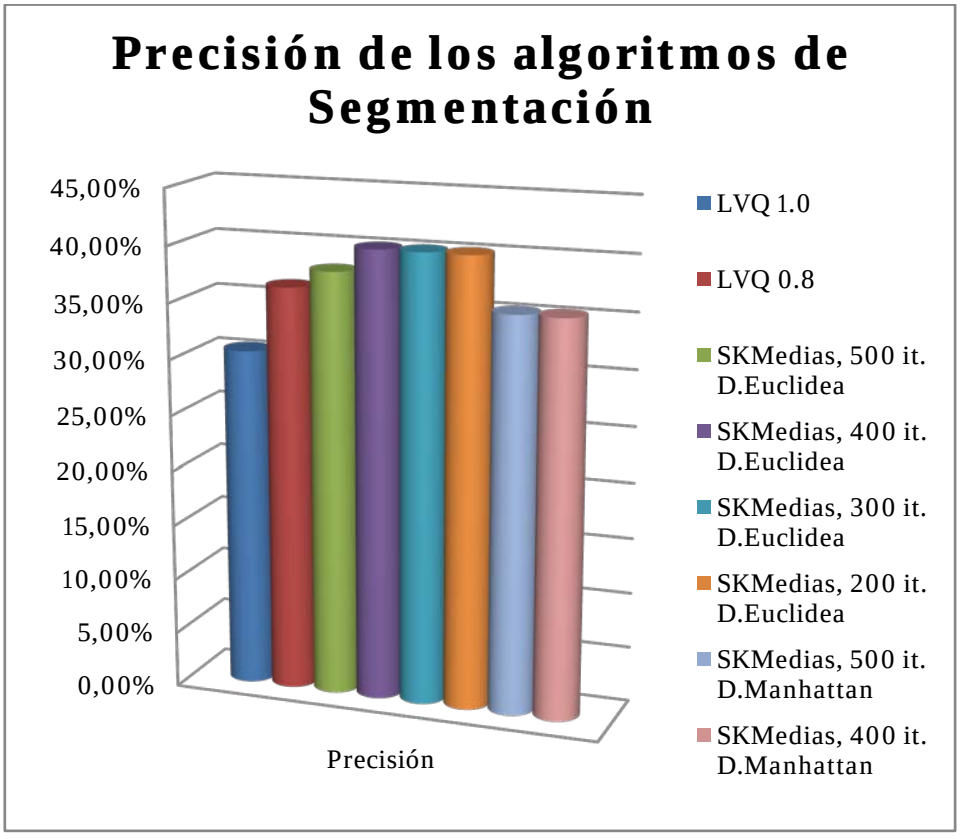


FIGURA 54 - Resultados de la Segmentación para todas las Clases

Como se puede apreciar en la imagen, SKMedias es el algoritmo que mejor agrupa en relación a la clase.

4.5 Evaluación de los Modelos

En base a los resultados de la segmentación para los algoritmos Canopy, LVQ y SKMedias, cuyas tablas recuperamos (véase Tabla 12, Tabla 13, Tabla 14, Tabla 16).

Se observa que las clases RE, DS y TR son las clases en las que identificar un patrón es más sencillo.

Estos resultados obedecen a que el número de instancias de esas ondas es mucho mayor en comparación al resto.

En cambio, si resulta relevante los clústeres que se generan y no se asignan a ninguna onda en LVQ y SKMedias (véase Tablas 15, 17).

En función, del algoritmo que se emplee se dejan sin agrupar las ondas VT, ND, NC, TO y NULL, coincidiendo en ambos casos las ondas VT y ND. Este hecho es relevante porque **no está ligado directamente al número de instancias**.

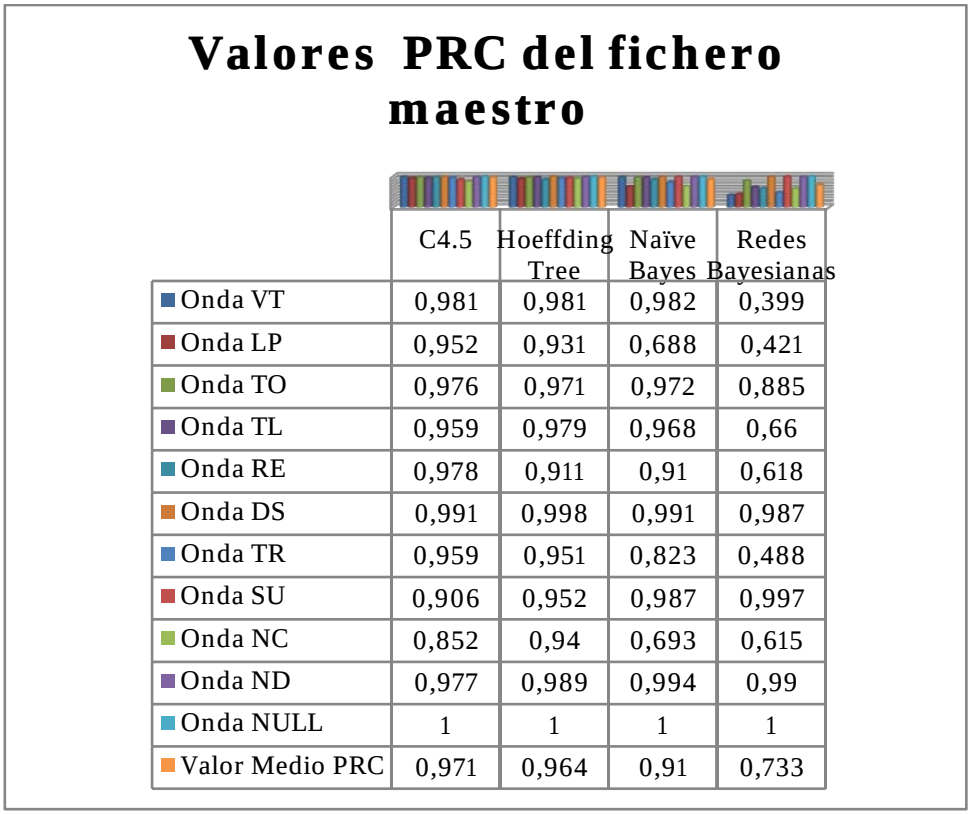
El número de instancias asociadas a la onda SU es inferior al de la onda VT y sin embargo, sí se segmenta en todos los casos. Las instancias asociadas a las ondas NULL y NC son dos de los subconjuntos más grandes y sin embargo, los algoritmos en el caso de LVQ son incapaces de detectar un patrón entre los datos.

Onda	Número de Instancias	Máximo de Instancias/Clase	Mínimo de Instancias clase
VT	4200	LP (57800)	SU (3614)
TO	5900		
NC	23439		
ND	7500		
NULL	23001		

Tabla 19 - Instancias de Clases No Segmentadas

Este hecho sin embargo, no se refleja en los clasificadores elaborados a partir del mismo fichero incluso antes de ser balanceados.

Muestran unos valores muy superiores al 0,5 para los clasificadores C4.5, Hoeffding Tree y Naïve Bayes; si bien, las Redes Bayesianas demuestran ser más sensibles y los valores de las ondas VT, LP, TL, RE, TR y NC oscilan entre 0,399 y 0,66.



	C4.5	Hoeffding Tree	Naïve Bayes	Redes Bayesianas
Onda VT	0,981	0,981	0,982	0,399
Onda LP	0,952	0,931	0,688	0,421
Onda TO	0,976	0,971	0,972	0,885
Onda TL	0,959	0,979	0,968	0,66
Onda RE	0,978	0,911	0,91	0,618
Onda DS	0,991	0,998	0,991	0,987
Onda TR	0,959	0,951	0,823	0,488
Onda SU	0,906	0,952	0,987	0,997
Onda NC	0,852	0,94	0,693	0,615
Onda ND	0,977	0,989	0,994	0,99
Onda NULL	1	1	1	1
Valor Medio PRC	0,971	0,964	0,91	0,733

Tabla 20 - Examen Valores PRC (Fichero Maestro)

En conclusión, el balanceo de datos ha servido para mejorar los ya de por sí, buenos resultados en los clasificadores y no hubiera sido necesario realizarlo.

El dominio es suficientemente explicativo y consigue por sí mismo, sin apenas ayuda ni manipulación de los datos realizar clasificadores de manera sobresaliente.

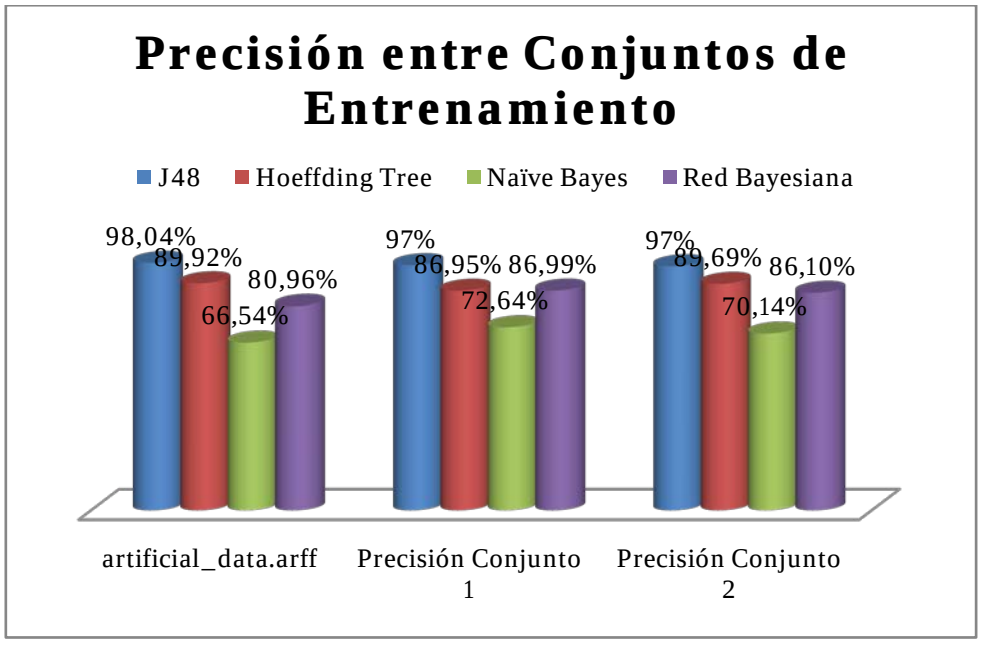


Tabla 21 - Evaluación final de los Clasificadores

5. Conclusiones y trabajos futuros

El trabajo realizado muestra las técnicas de Minería de Datos y el campo de la sismología como un terreno fértil en el que poder trabajar. Elaborar primero una clasificación y posteriormente una segmentación para contrastar los resultados, me ha resultado interesante. Si bien es cierto que estas tareas no suelen estar relacionadas ahondar en los resultados obtenidos en una de ellas apoyándose en la otra es cuánto menos interesante.

5.1 Conclusiones del estudio

Al iniciar el proyecto, se tomó la decisión de elaborar modelos de clasificación en base a los datos facilitados por el Instituto de Geología y Minas de Colombia y estudiar si su elaboración podría llegar a ser lo suficientemente precisa para poder realizar en el futuro esta tarea manual de forma automática.

Los primeros pasos consistieron en adquirir los datos y pre-procesarlos para poder ser empleados por Weka. Este pre-procesamiento fue un contratiempo, que si bien se solventó con éxito incidió gravemente en la planificación (véase Anexo C).

Con los ficheros preparados se procedió a elaborar un fichero maestro que contuviera el máximo de ondas posibles, en total se recuperaron 11 ondas de las 18 ondas a recuperar. Posteriormente se crearon otros dos conjuntos de datos, Conjunto 1 y Conjunto 2 fruto de análisis estadísticos practicados sobre el fichero maestro.

En todos los casos, los resultados que arrojaban los algoritmos de clasificación sobre los conjuntos rondaban una tasa de acierto próxima al 99% en tres de los cuatro algoritmos (C4.5, Hoeffding Tree, Naïve Bayes) y de 70% en el cuarto (Redes Bayesianas)).

Estos resultados resultaban sospechosos, por lo que se decidió realizar un estudio más profundo de los datos realizando experimentos de segmentación sobre el fichero maestro.

Estos experimentos hallarían patrones relevantes y desconocidos en los datos en caso de haberlos y ofrecerían evidencias con las que respaldar los resultados obtenidos en la clasificación y medir la influencia del sesgo aplicado para balancear los conjuntos Conjunto 1 y Conjunto 2.

A medida que se realizaban los experimentos de segmentación se observaba cierta lógica en sus resultados: los clústeres asociados a ondas de los primeros experimentos que contaban con apenas 3 ó 5 clústeres (un tercio y la mitad de las clases aproximadamente) aparecían asociados a las mismas ondas en posteriores experimentos con 11 clústeres (número total de clústeres).

La repetición de estos clústeres en todos los experimentos se debía simplemente a que el número de instancias por clase era muy superior al del resto. Sin embargo, se realizó un hallazgo interesante, segmentos a los que no se les había asociado clase no pertenecían a las clases menos numerosas, si no a clases poco y muy numerosos, y lo

más interesante es, que este hecho apenas había tenido impacto en los clasificadores, ni siquiera en aquellos cuyo conjunto de entrenamiento era el fichero inicial, sin balanceo de las clases.

Los experimentos realizados permiten concluir que es posible a partir de un conjunto de entrenamiento del que se prescinda del atributo fecha, elaborar un clasificador fiable y atemporal de 11 de los 18 tipos de ondas que recogen los sismógrafos en torno al volcán Puracé.

En el plano personal, este proyecto amplió mis conocimientos sobre sismología, un área de mis intereses particulares y aprender acerca del análisis de datos, etapas, metodologías y límites de sus técnicas.

5.2 Trabajos futuros

Como ya se describió en el planteamiento del proyecto, este estudio simplemente explorar la posibilidad de elaborar un clasificador preciso de ondas sísmicas. Una posible línea futura de este trabajo, sería elaborarlo. Para ello se necesitaría de un conjunto de entrenamiento suficientemente grande y que contuviera los 18 tipos de ondas.

Otra línea de trabajo sería la de repetir experimentos de clasificación ya realizados empleando herramientas de minería de datos que permitieran generalizar los resultado. Realizar experimentos de segmentación también resultaría interesante.

Hasta ahora, los estudios referidos a volcanes son locales, es decir, es habitual que se centren en la actividad de un único volcán. Las técnicas de Minería de Datos permitirían probar una idea que incluyera varios volcanes simultáneamente. Los estudios de segmentación podrían determinar por ejemplo, zonas en las que de ocurrir un terremoto la probabilidad de que se halle un epicentro es máxima.

Por supuesto, todos estos trabajos tienen como última finalidad una mayor comprensión del dominio para poder prevenir con éxito erupciones y terremotos.

6. Planificación

En esta sección se expone la planificación del proyecto, desde la planificación inicial, elaborada el mismo día en que se inició para establecer las pautas e hitos a seguir, a la planificación final corregida y ajustada al desarrollo real del proyecto.

En la primera planificación sólo se recogieron las fases principales del proyecto, dado que no se conocía su alcance total. Se planificaron 7 fases cuyo diagrama Gantt correspondiente se puede consultar en el Anexo C (Tabla 24: Planificación inicial). Las fases son las siguientes:

- Fase inicial: Que abarca desde el principio del proyecto hasta establecer los objetivos, adquirir los datos y entenderlos. Duración de 10 días.

- Investigación: En esta etapa se busca y elabora documentación relacionada con el ámbito del proyecto. Duración de 17 días.

- Experimentación: Desarrollo de las pruebas relativas a la solución planteada para el problema. Duración de 20 días.

- Elaboración de la memoria: Elaboración de este documento, que refleja todos los aspectos abordados en su realización. Duración de 7 días.

- Contratiempos: Abarca toda clase de contingencias, desde pérdida de un documento hasta la ampliación de pruebas en la Experimentación y su posterior inclusión en el documento. Duración de 9 días.

- Revisión General: En esta observa se revisa que el documento cumpla con los requisitos señalados en la matriz de evaluación y las pautas indicadas por el tutor. Duración de 4 días.

- Entrega: Entrega final del documento. Duración de 1 día.

- Revisión general: La Revisión General amplió su duración a 11 días. La planificación inicial preveía una duración total del proyecto de 60 días más 9 días de contratiempos, lo que hacen un total de 69 días. La previsión de trabajo se fijó en 8 horas diarias, lo que hacen un total de 552 horas de trabajo incluyendo contratiempos. Sin embargo, la duración inicial del proyecto se vio alterada al tener que elaborar un programa en Java que preparara elaborara ficheros para realizar la experimentación a partir de los datos iniciales y realizar experimentos de segmentación no previstos al inicio del proyecto. Las etapas de experimentación y contratiempos se vieron afectadas. La experimentación amplió su duración a 39 días y la de contratiempos a 67 días, ya que ambas etapas se solaparon en el tiempo junto a la elaboración de la memoria. Las partes de mayor duración han sido la elaboración de la memoria con 83 días, ya que prácticamente se inició en el momento en que lo hizo la etapa de investigación y no finalizó hasta producirse la entrega, la fase de experimentación, con 39 días, y la de contratiempos con 67 días y que comprende las tareas de programación, nuevos clasificación y segmentación. La planificación final refleja una duración total del proyecto de 107 días, con una media de 6 horas trabajadas en las primeras fases del desarrollo (30 días) y más de 8 horas trabajadas por día en el resto de fases, lo que hace un total de 806 horas de trabajo totales. La planificación final completa se puede consultar en el Anexo E (Tabla 26: Planificación final).

7. Entorno socio-económico: presupuesto

A continuación se presenta una estimación de costes de desarrollo en términos de personal, equipamiento y licencias de software. Para el cálculo de costes imputables para equipamiento y licencias, se ha utilizado un periodo redondeado de duración del proyecto de 4 meses, además del tiempo de amortización desde la fecha de adquisición del equipo/software.

- Autora: María de las Mercedes Crespo Jiménez
- Departamento: Departamento de Informática
- Proyecto: Clasificación de ondas sísmicas con técnicas de minería de datos
- Desglose:

PERSONAL			
Categoría	Salario	Horas trabajadas	Coste total
Investigadora	**20,00€/h**	**450h**	**9.000,00€**
			Total (€): 9.000,00 €

EQUIPAMIENTO					
Descripción	**Coste**	**Duración**	**Uso dedicado al proyecto**	**Período de amortización**	**Coste imputable**
PC Portátil Toshiba 500-B Satellite	**800€**	**4 meses**	**100%**	**36 meses**	**300€**
					Total (€): 300 €

SOFTWARE					
Descripción	**Coste**	**Duración**	**Uso dedicado al proyecto**	**Período de amortización**	**Coste imputable**
Ubuntu 17.04	**0 €**	**4 meses**	**100%**	**-**	**0 €**
Microsoft Office 2010 Professional	**300€**	**4 meses**	**30%**	**36 meses**	**30,50€**
Eclipse	**0 €**	**4 meses**	**40%**	**-**	**0 €**
Notepad ++	**0 €**	**4 meses**	**50%**	**-**	**0 €**
WEKA	**0 €**	**4 meses**	**100%**	**-**	**0 €**
Microsoft Project	**0 €**	**4 meses**	**100%**	**-**	**0 €**
					Total (€): 30,50 €

Tabla 22 - Costes del proyecto

RESUMEN	
Concepto	**Coste**
Personal	9.000,00 €
Equipamiento	30,50 €
Software	0 €
Total sin IVA	9030,50 €
IVA (21%)	1896,405€
Total	10,926,905 €

Tabla 23 - Presupuesto del Proyecto

Bibliografía

Alpala J., Alpala R., Battaglia M. «Monitoring remote volcanoes: The 2010-2012 unrest at Sotará volcano (Colombia).» *Journal of Volcanology and Geothermal Research*, 2017.

Berzal, F. «DECSAI (Departamento de Ciencias de la Computación e IA) de la Universidad de Granada.» 2017. http://elvex.ugr.es/decsai/intelligent/slides/dm/D2%20Association.pdf.

Blázquez García, Raúl. *Estudio de técnicas de Análisis de Datos aplicadas al Seguimiento de Imágenes*. (Trabajo de Fin de Grado, Universidad Carlos III), 2004.

Chatfield, Chris. *The Gallery of Natural Phenomena.* 2010. http://www.phenomena.org.uk/earthquakes/earthquakes/lisbon.html.

Cortina, Víctor Galán. *Aplicación de la Metodología CRISP-DM a un Proyecto de Minería de Datos en el Entorno Universitario.* Trabajo de Fin de Grado (Universidad Carlos III de Madrid), 2015.

Datos, Agencia Española de Protección de. 1999. http://www.agpd.es.

Davison, Charles. *The Founders of Seismology.* New York: Arno Press, 1978.

Fayyad U. M, G Piatetsky-Saphiro, P.Smyth, and R.Uthurusamy. *Advances in Knowledge Discovery and Data Mining.* 1996.

Fernández Rebollo, F. *Transparencias de la Asignatura de Aprendizaje Automático.* Universidad Carlos III de Madrid, 2015.

Ferrer–Troyano, Francisco J. y Aguilar–Ruiz, Jesús S. *Minería de Data Streams: Conceptos y principales Técnicas.* Sevilla: Universidad de Sevilla, 2005.

García Jiménez, Beatriz. *Estudio del Fracaso Escolar mediante Técnicas de Minería de Datos.* Trabajo de Fin de Grado (Universidad Carlos III), 2005.

Hssina B., Merbouha A.,Ezzikouri H., Erritali M. «A comparative study of decision tree ID3 and C4.5.» *(IJACSA) International Journal of Advanced Computer Science and Applications,Special Issue on Advances in Vehicular Ad Hoc Networking and Applications* 13-19.

Kirby, R., Hall, M. *weka.sourceforge.net.* http://weka.sourceforge.net/doc.dev/weka/classifiers/trees/HoeffdingTree.html (último acceso: 2017).

Mcnutt, Scott R. «Volcanic Tremor Amplitude Correlated With Eruption Explosivity and its Potential Use in determining Ash Hazards to Aviation.» *U.S. Geological Survey Bulletin 2047*, 1994.

McNutt, Stephen R. «Volcanic Tremor .» En *Encyclopedia of Earth System Science, Volume 4*. California: Academic Press, Inc., 1992.

Mena, Jesús. *Machine Learning Forensics for Law Enforcement, Security, and Intelligence.* Boca Raton: FL: CRC Press (Taylor & Francis Group), 2011.

Minería, Instituto Colombiano de Geología y. «Transaparencias de Sismología.» Instituto Colombiano de Geología y Minería.

National Institute of Geophysics and Volcanology. *Workshop on 25 years advancing volcano seismology in a wider volcanological context (25th European Seismological Commission working group volcano Seismology).* Stromboli, Italia: Miscellanea INGV, 2016.

Piatetsky-Shapiro, Gregory. «Knowledge Discovery in Real Databases: A Report on the IJCAI-89 Workshop.» *AI Magazine, Volume 11, Number 5*, January 1991: 68-70.

Quinnlan, J.R. 1985. http://hunch.net/~coms-4771/quinlan.pdf.

Rodríguez Montequín, Mª Teresa, J. Valeriano Álvarez Cabal, José Manuel Mesa Fernández, y Adolfo González Valdés. *Metodologías para la Realización de Proyectos de Data Mining.* Oviedo: Departamento de Matemática Aplicada (Universidad De Oviedo).

Rojas, Dr. Oldemar Rodríguez. *Metodología para el Desarrollo de Proyectos de Minería de Datos CRISP-DM.* 2010.

Saito T, Rehmsmeier M. *The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.* 2015. http://journals.plos.org/plosone/article?id=10.1371/journal.pone.0118432.

SIGKDD. *KDD.* 2016. http://www.kdd.org/curriculum/index.html.

StackExchange. 2014. https://stats.stackexchange.com/questions/90779/area-under-the-roc-curve-or-area-under-the-pr-curve-for-imbalanced-data (último acceso: Septiembre de 2015).

Telenchana E., Bernard B., Hidalgo S., Beate B. «Modelo evolutivo del volcán Chiles.» 2017. https://www.researchgate.net/publication/317033772.

Turban, E., Sharda, R. y Dursum, D. *Decission Support and Business Intelligence Systems.* Prentice Hall, 2011.

Unglert K., Jellinek, A.M. *Journal of Volcanology and Geothermal Research*, 2017.

Valls, J.M. *Clasificación (LVQ).* Leganés: Transparencias de la Asignatura Redes de Neuronas. Aprendizaje no Supervisado (Universidad Carlos III de Madrid), 2017.

Witten, Ian H., Frank, E., Hall, Mark A., J. Pal, Christopher. *Data Mining Practical Machine Learning Tools and Techniques, 4th Edition.* Morgan Kaufmann, 2016.

ANEXO A: English Summary

Abstract

Given the seismic activity increase of the last 20 years, it's mandatory to research in this field from a different point view. Technology has updated the tools and mechanisms that seismologists were already using but their research methods haven't been updated yet. They're still working with the same information they had in the last 80 years and their experiments are taking as long as they did in the past retrieving almost the same information.

Machine learning techniques could improve their research by analyzing huge amounts of data and extracting new knowledge. Data mining techniques are already being used in many different fields, like medicine, car insurance or publicity successfully. Why not in seismic activity?

In this project, segmentations and classification techniques will be used in order to extract knowledge of the domain and explore the possibility of building accurate seismic waves classifiers and extract new information from the domain.

Introduction

Seismology is the scientific study of earthquakes and the propagation of elastic waves through the Earth or through other planet-like bodies. The purposes of the seismology are:

1) Know the Earth's structure.
2) Study the causes of seismic movements.
3) Prevent damages.
4) Study phenomena related with earthquakes, like tsunamis or volcanoes' dynamics.

This Project is focused on the last point. Its aim is make a study that guarantees the feasibility of a seismic wave's classifier using the information of Purace's volcano seismographs in the period of 1-7-2015 to 1-7-2016.

Earthquakes

An earthquake is a seismic movement whose origin is in a specific point of a fault called **focus**. The name of this point in the ground is **epicenter**.

Seismic activity has its origin in natural causes (plaques tectonic) and human activity likes explosions underground or nuclear bombs.

Tremors of seismic waves are captured by **seismographs** which are studied after by experts in order to classify the waves that have registered, localize the epicenter or the focus or forecast.

Data Mining

Data Mining (*DM*) is strongly related with KDD (*Knowledge Discovery in Databases*). KDD is the process of discovering useful knowledge from a huge collection of data including data preparation and selection, data cleansing, incorporating prior knowledge on data sets and interpreting accurate solutions from the observed results. Data Mining is just a part of KDD's process.

Data mining results need to be interpreted by an expert. Without interpretation of the results, there's no answer to problem and without an answer the problem, the question remains.

Data mining techniques allow to:

- **Predict tendencies and behaviors:** find out why two events are related may help for example, to discover how to improve the effects of a medicine or the best moment to make a purchase or a sale.

- **Discover unknown models and patterns:** often causality and casualty are mistaken. Data mining techniques help to identify if two co-occurring or consecutive events are or not correlated.

- **Analyze huge databases:** eliminate rows and columns of a database may lead to eliminate relevant information or apply bias. Data mining techniques can analyze a whole database without reducing its dimensionality.

Major KDD application areas include:

- **Customer relationship management (CRM)**: its goal is to create one-on-one relationships with customers by developing an understanding of their needs and wants. Customer's relationships management are usually long term relationships and so, as businesses build relationships with their customers tons of data are collected. These data allow companies to elaborate customer profiles, understand the roots causes of customer attrition in order to improve customer retention or discover time-variant associations between products and services.

- **Banking**: data mining perform the following tasks, automating the loan application process by predicting the most probable defaulters, detecting fraudulent transactions, optimizing the cash return by forecasting the cash flow on banking entities and like in CRM, data mining can identify ways to maximize customer value by selling them products and services that they are most likely to buy.

- **Retailing and logistics**: DM is used to predict accurate sales volumes at specific retail locations in order to determine inventory levels and forecast consumption levels of different product types to optimize logistics and hence maximize sales.

- **Insurance:** data mining technique are used to for example, predict which customers are prone to buy new policies with special features or identify and prevent incorrect claim payments and fraudulent activities.

- **Entertainment industry:** DM techniques are used to analyze viewer data to decide how to schedule programs and advertisement or predict the financial success of movies before they are produced to make investment decisions and to optimize the returns.

- **Sports:** game patterns can be identified by analyzing basketball, tennis or football videos. These patterns could be used to coach teams or design better plays.

- **Medicine:** DM is used as a complement to traditional medical research. Data mining analyses can identify novel patterns in medical data, predict success rates of highly risky operations like transplants or discover relationships between symptoms and illnesses.

Data Analysis

Data are the lowest level of abstraction from which information and knowledge are derived. Data refers to a collection of facts usually obtained as the result of experiences, observations, or experiments.

First step to perform a data analysis is getting the data. Data are usually retrieved from databases but it's also possible to acquire them from other sources, like real data streams.

Following steps are understanding business and data in order to establish objectives and evaluate if the given dataset is representative and enough accurate by performing statistical techniques.

Next step is get data prepared by selecting and scrubbing the original data.

Data selection is conducted with the purpose of eliminating irrelevant data or replacing specific attributes with a combination of some other attributes (for example, the mean of some values may replace those values in the dataset).

While scrubbing outliers and absent data are detected and one of the following tasks is performed:

- **Ignore:** only if algorithms does not perform well if they are outliers or absent data in the dataset.

- **Replace value:** outlier's value is replaced by a value calculated using attribute's value of the dataset.

- **Discretize:** data get homogenized. Outliers get labeled or grouped around a specific value.

- **Eliminate o replace a column:** this option is not very advisable; the risk of losing relevant information is high.

- **Filter a row:** alter a row directly will introduce a bias but if the dataset is big enough it may not have any influence.

- The best way to treat **absent data** is waiting until they'll be available but if this possibility does not exist, it's pretty convenient to find out why they are absent, typically due to one of the following reasons:
  - **It does not exist:** for external reasons a value has not been already calculated or is available.
  - **Its absence is meaningful.**
  - **Incomplete information**: if data are retrieved from several it's possible that some values remain empty.

Machine learning

Building any DM model requires extract information, identify and learning patterns from data. This learning process can be supervised or unsupervised.

Supervised learning: every row of the dataset has its corresponding target. Supervised learning algorithms seek a function from inputs to the respective targets. Its name comes from the algorithm learning process of the training dataset; it can be thought of as a teacher supervising the learning process.

If the targets are expressed in some classes, it is a classification problem. Alternatively, if the target space is continuous, it is called regression problem. Learning stops when the algorithm achieves an acceptable level of performance.

Unsupervised learning: Contrary as supervised learning, inputs has no label associated. It's called unsupervised because unlike supervised learning there is no teacher saying what's right or wrong.

Unsupervised learning allows finding the underlying structure or distribution in the data in order to learn more about the data.

How Data Mining Works

From data, data mining builds models to identify patterns among the attributes presented in the dataset. Models are the mathematical representations that identify the patterns among the attributes of the objects described in the dataset. These patterns may be explanatory or predictive; four major types of patterns are identified:

Prediction

Taking into account experiences, opinions and other relevant information it's possible to get an idea of events that may occur in the future if circumstances are similar or repeated.

Two main tasks are performed: classification and regression. In classification or supervised induction data are subdivided into a specific number of categories, with a nominal attribute; regression does this subdivision over data whose classification is numerical.

Several techniques can be used; some of them are explicative like decision trees while others focus on the results like neural networks.

Association

Identify the most common co-occurring grouping of things. It's a popular technique, very useful and effective to discover interesting relationships among variables in large databases.

Clustering

This technique partitions a collection of things into segments whose member share similar characteristics, but unlike in classification class labels are unknown. Clusters are determined using a heuristic-type algorithm and so, their results are strongly correlated with the algorithm.

That's why, it's recommended to use several algorithms and extract knowledge from most repeated clusters among them. The goal of clustering is to create segments whose members within each group have maximum similarity and the member across them minimum similarity.

Data Mining Tool: WEKA

In this Project all the experimentation will be done using the Data Mining tool WEKA. This is an open-source tool designed and created by the University of Waikato (New Zealand).

This tool can be used with Ubuntu, MAC and Windows Operating Systems. Due to the fact, I'm currently using Ubuntu and I have got some experience using this tool, I think it's appropriate to do this Project.

WEKA will analyze all the dependencies among the attributes of my dataset and will elaborate the classifiers and output the results of clustering experiments.

WEKA's version for this project is 3.9.1. In addition of its own functionalities, some externals packages were added to run incremental learning instead of only batch learning.

CRISP-DM Metodology

Experimentation will follow CRISP-DM (*Cross Industry Standard Process for Data Mining*) process model whose structure has six steps:

Business Understanding

This initial phase focuses on understanding the project objectives and requirements from a business perspective and converts this knowledge into a data mining problem definition.

It's mandatory a fully understanding of the data and the objectives, otherwise the results may not be useful to solve a problem in the real world.

Data Understanding

This phase starts with an initial data collection and proceeds with analysis that may identify data quality problems, help discover first insights into the data, or detect interesting subsets to form hypotheses for hidden information.

This phase is usually the one which requires most time and effort and it's pretty important to pay attention to it. If data are not well understood, their results wouldn't be understood or even worse, misunderstood and lead to wrong solutions.

Data Preparation

This phase covers all activities to construct the final dataset. Data preparation tasks are likely to be performed multiple times, and they don't follow any strict order. Some of these tasks are elaborate tables of date, chose the best visualization data techniques, select and transform attributes or cleaning the dataset.

This phase is highly related with the Modeling phase. Some techniques have specific requirements on the form of data and they must be pre-processed in a specific way that may not be useful for other techniques. Therefore, stepping back from the Modeling phase is often needed.

Modeling

In this phase, various modeling techniques are selected and applied depending on the Data Mining task they are performing. Typically, there are several techniques for the same data mining problem type so they are usually chosen according to the following criteria:

- It's appropriate to solve the problem.
- Data are correct.
- It respects problem's requirements.
- Models are generated quickly.
- The technique is known, well-researched and documented.

Evaluation

At this point, a single model (or several) have been created and appears to have high quality but it must be proved. It is important to evaluate thoroughly the model, and review the steps executed to construct the model, to be certain it properly achieves the objectives.

At the end of this phase, a decision on the use of the data mining results should be reached.

Deployment

This phase is the phase in which all the process is documented and the obtained knowledge of model is presented in a clear and useful way. This project does not contain this final step, but a summary of the results of the Experimentation and the elaboration of the Project.

Algorithms

In this project classification and clustering algorithms will run on the data. Classification algorithms are C4.5, Hoeffding Tree, Naïve Bayes and Bayesian Nets. Clustering algorithms are only three: Canopy, LVQ and Simple K-Means.

Classification algorithms

C4.5

C4.5 is an algorithm used to generate a decision tree developed by Ross Quinlan based on Quinlan's ID3 decision tree.

C4.5 builds decision trees from a set of training data in the same way as ID3, using the concept of information entropy but has some improvements:

- Handles both continuous and discrete attributes.
- Handles training data with missing attribute.
- Handles attributes with differing costs.
- Prunes trees after creation

At each node of the tree, C4.5 chooses the attribute of the data that most effectively splits its set of samples into subsets enriched in one class or the other. The splitting criterion is the normalized information gain (difference in entropy). The attribute with the highest normalized information gain is chosen to make the decision.

Hoeffding Tree

Hoeffding tree (VFDT) is an incremental tree induction based on ID3. Its algorithm is capable of learning from massive data streams, assuming that the distribution generating examples does not change over time. It can run in either batch or incremental mode.

Its name comes from Hoeffding bound which quantifies the number of observations (in our case, examples) needed to estimate some statistics within a prescribed precision (in our case, the goodness of an attribute).

They're also pretty effective thanks to its splitting criteria. A small training set's sample can often be enough to choose optimal splitting attributes. Its output can also be compared with other decision tree learners' output. Using the Hoeffding bound one can show that its output is asymptotically nearly identical to that of a non-incremental learner if using infinitely many examples.

Baye's law

When it comes to forecasting, Baye's rule is used to describe the probability of an event, based on prior knowledge of conditions that might be related to the event.

Bayes' theorem has the following equation where:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

- $P(A)$, $P(B)$ are the probabilities of observing and without regard to each other.
- $P(B|A)$ is a conditional probability, is the probability of observing event given that is true.

- $P(A|B)$ is the probability of observing event given that is true.

Naïve Bayes and Bayes network are algorithms run in WEKA both based on Baye's rule.

Naïve Bayes generates a fixed Bayes network structure with arrows from the class variable to each of the attribute variables using numeric estimators whose values are chosen based on analysis of the training data.

Bayes networks look pretty much like a Naïve Bayes algorithm's output. It looks exactly like a Naïve Bayes output but it has random instances that Bayes Network learning algorithm has created using various search algorithms and quality measures to build the model.

Segmentation algorithms

Canopy

This unsupervised algorithm clusters the algorithm into proximity regions, canopies. Two heuristics are needed to calculate how many canopies will be created, T1 and their size, T2, with T2 < T1. These values can be generated randomly or chosen.

Its execution has two stages: Su ejecución se divide en dos fases en la que primero se crean los clústeres y después se asignan instancias:

- **Stage 1**: canopies are created and one aleatory example of the training set is assigned to each one. Then, the distance between canopies and the rest of instances is calculated. If one of these distances is bigger than T2, another canopy will be created or ignored.
- **Stage 2:** using T1 instances are associated to a canopy. Sometimes, a canopy belongs to more than just one cluster.

K-Means

K-means clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. Those prototypes are chosen randomly in order to not create bias that may modify the way clusters are created but the number of clusters *k* can also be chosen.

Once the prototypes are defined, instances start to cluster around them taking into account, the value of its mean and the distance between the object and the cluster.

With each new example, prototypes and instances values of clusters are re-calculated until all value's object are stable.

LVQ (Learning Vector Quantization)

LVQ is the supervised version of Self Organising Maps (SOM).

It has no neighborhood, and only the winner cell is modified. SOM modifies the whole neighborhood while LVQ modify the winner cell. Cell's direction change approaching or moving away from a single cluster while in SOM the final direction of the cell while in SOM the approach and distancing involves all the clusters. Learning rate decreases in time while in SOM it's constant throughout all the experimentation.

First algorithm's step is distributing prototypes randomly across the entry space. The number of prototypes that clusters may contain can be defined as a single random number to all of them or as a future measure related to the number of instances that clusters will contain.

Experimentation

The Colombian Geological Survey gave to University Carlos III the seismic activity and its analysis of Puracé volcano's during 1-7-2015 to 31-12-2016.

Seismic activity was recorded and analyzed daily. A record was a .txt with the seismic activity of a particular day. This seismic activity was the voltage of the seismic wave measured by Aguas Blancas, Cocuy, Condor, Lavas Rojas, Pilimbala and Shaka seismographs. This voltage was measured in components E, N, Z for each seismograph.

Each .txt has an analysis kept in a .csv with the same name. Each analysis contained 9 attributes including the seismic wave associated to each seismic activity.

In order to make the experimentation, new files associating .txt's and .csv's were created. So the new txt's contained their original information but also the wave they were related to.

This operation was only made to 1-8-2015 - 31-12-2015 files. August 2015 data contained only the seismograph's records while all the information inside 2016's file were the analysis of volcanic monitoring.

Once these files were created, it was time to prepare the master file. This master file master must contain the 18 types of seismic waves and must be small enough to run in WEKA and big enough to be representative.

Only 11 types of seismic waves were retrieved and a file of 32,9MB was created.

Data Analysis

The file master was ready. It had only 19 attributes because they were no records of Pilimbala's seismograph.

Four data analyses were performed to reduce the dimensionality of the set. Two sets were created; one with 9 attributes the other with 7.

In any case, these four techniques show the importance of the attribute to the dataset but each one in its particular. Set with 9 attributes was called Set 1, set with 7, Set 2.

Set 1 was created using a technique of evaluation of subsets for classifiers which directly point out the relevant attributes: 100% if the attribute is relevant, 0% if it's not.

Set 2 was created mixing the result of the three others techniques. Those techniques elaborate rankings of the attributes. In order to reduce the dimensionality of the set, only the upper half of the attributes were analyzed; if an attribute was in the upper half and appeared in the three studies, then, it was chosen to join the other attributes of Set 2. These techniques were: correlation analysis among attributes and gain ratio information of the attributes respect their class, in this case the value of the specific wave they were identifying.

In addition, Set 1 and Set 2 were balanced while the master file remained with the value of its classes unbalanced.

Classification

C4.5, Hoeffding Tree, Naïve Bayes and Bayes Net algorithms were used to create the classifiers. All of them gave great results no matter which training set was being used (file master, Set 1, Set 2): C4.5, Hoeffding Tree and Naïve Bayes had accuracy close to 99% and Bayes Net to 70%.

Given these high results, some clustering experiments were done with the purpose of getting some new information and justify classifiers.

Clustering

Clustering results were made using only the file master. Canopy, LVQ and SKMeans were used to explore the data. These experiments allow identifying which seismic waves had the biggest clusters and why.

Surprisingly, they were RE, DS and TR, the waves that had the biggest number of instances. This fact may be quite obvious. Next discover is far more interesting.

LVQ and SKMeans experiments identify two waves that never cluster: VT and ND. They never cluster in both algorithms and this fact has no relation with the number of instances per class. SU class is far smaller than VT class. Nevertheless they always cluster for every LVQ and SKMeans experiment.

Some other big classes didn't cluster too. For example, LVQ couldn't find a pattern for NULL and NC though they are two of the biggest classes of the whole dataset.

Result's evaluation

Segmentation results show how important the number of instances per class is to perform an accurate and useful analysis but also how sensitive can some algorithms be even when the dataset is strongly unbalanced.

Segmentation experiments identify the waves with the biggest clusters as the one with the biggest amount of instances but they also identify waves whose patterns are not clear enough. More experiments are necessary to conclude the bias impact of the result.

Nevertheless, classifiers whose data belong to the master file output in an extremely good way. Their PRC results are higher than 0, 5 for each class of each algorithm in exception of Bayes Nets which detected PRC values between 0,399 and 0,66 for VT, LP, TL, RE, TR and NC waves.

No balance among classes and scrubbing was needed. Master file results were already good.

Next experiments must include the same name of instances per wave and attention should be focus in clustering experiments because the information we can obtain of a domain is pretty interesting, with a more heterogeneous and bigger training set, who knows what could be find?

Conclusions

This project has taught that good results are not quite enough to assure accuracy and that the knowledge of a domain is capital to interpret the results and design experiments.

ANEXO B: Capturas de los Experimentos

Fichero Maestro

C4.5

```
Time taken to build model: 41.79 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      264817               98.0357 %
Incorrectly Classified Instances      5306                1.9643 %
Kappa statistic                          0.9767
Mean absolute error                      0.0041
Root mean squared error                  0.0583
Relative absolute error                  2.6388 %
Root relative squared error             21.0467 %
Total Number of Instances           270123
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,985     0,000     0,990       0,985   0,987      0,987    0,996     0,981     00000
               0,967     0,007     0,967       0,967   0,967      0,960    0,984     0,952     00001
               0,985     0,000     0,985       0,985   0,985      0,985    0,996     0,976     00100
               0,976     0,002     0,975       0,976   0,976      0,974    0,991     0,959     00101
               0,985     0,005     0,983       0,985   0,984      0,980    0,993     0,978     00110
               0,994     0,002     0,993       0,994   0,993      0,992    0,998     0,991     00111
               0,973     0,005     0,974       0,973   0,973      0,968    0,989     0,959     01000
               0,928     0,001     0,954       0,928   0,941      0,940    0,973     0,906     01010
               0,888     0,001     0,919       0,888   0,903      0,902    0,967     0,852     01011
               0,988     0,001     0,982       0,988   0,985      0,984    0,996     0,977     10001
               1,000     0,000     1,000       1,000   1,000      1,000    1,000     1,000     11111
Weighted Avg.  0,980     0,004     0,980       0,980   0,980      0,977    0,992     0,971
```

```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j     k   <-- classified as
  4135    13    10     3    38     0     0     1     0     0     0 |     a = 00000
    16 45919    39   121   306   348   618    71    38     9     0 |     b = 00001
     2    44  5813     0     1     0    39     1     0     0     0 |     c = 00100
     0   143     1 21091    73    29   225    24    14     0     0 |     d = 00101
    24   340     1    79 56951     0   146    32   199    28     0 |     e = 00110
     0   277     0    20     0 51801     1     1     0     0     0 |     f = 00111
     0   616    33   262   158     4 42293    19     6    93     0 |     g = 01000
     0   107     1    32    64     0    41  3354    11     4     0 |     h = 01010
     0    44     0    10   309     0    13    10  3053     0     0 |     i = 01011
     1     4     2     2    20     0    61     1     0  7409     0 |     j = 10001
     0     0     0     1     0     0     0     0     0     2 22998 |     k = 11111
```

FIGURA 55 - Salida C4.5 Fichero Maestro

Hoeffding Tree

```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j     k   <-- classified as
  3780    59    62     0   299     0     0     0     0     0     0 |     a = 00000
   187 33990   416  1181  2506  1334  7459   155   256     1     0 |     b = 00001
    45   111  5422     0     6     0   316     0     0     0     0 |     c = 00100
     0   395     0 20603    60    23   499     3    12     5     0 |     d = 00101
   180  1561    10    72 54876     0   476   134   450    40     1 |     e = 00110
     0   590     0     8     0 51487    15     0     0     0     0 |     f = 00111
     0  4082   756   541   466    51 36891    21   173   503     0 |     g = 01000
     0    41     1     4     5     0     6  3553     2     0     2 |     h = 01010
     0   378     0     1  1008     0    86     5  1961     0     0 |     i = 01011
     0     5     0     3    26     0   137     0     0  7328     1 |     j = 10001
    10     0     0     0     0     0     0     0     0     0 22991 |     k = 11111
```

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      242882               89.9153 %
Incorrectly Classified Instances     27241               10.0847 %
e attribute to use as the class                0.8807
Mean absolute error                            0.0229
Root mean squared error                        0.1177
Relative absolute error                       14.8851 %
Root relative squared error                   42.4886 %
Total Number of Instances           270123
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,900    0,002    0,900      0,900    0,900      0,898    0,998     0,966     00000
                 0,716    0,032    0,825      0,716    0,766      0,723    0,956     0,853     00001
                 0,919    0,005    0,813      0,919    0,863      0,861    0,997     0,944     00100
                 0,954    0,007    0,919      0,954    0,936      0,931    0,997     0,978     00101
                 0,949    0,021    0,926      0,949    0,938      0,921    0,993     0,972     00110
                 0,988    0,006    0,973      0,988    0,981      0,976    0,999     0,997     00111
                 0,848    0,040    0,804      0,848    0,826      0,791    0,978     0,895     01000
                 0,983    0,001    0,918      0,983    0,949      0,949    0,995     0,979     01010
                 0,570    0,003    0,687      0,570    0,623      0,622    0,982     0,682     01011
                 0,977    0,002    0,930      0,977    0,953      0,952    1,000     0,986     10001
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     11111
Weighted Avg.    0,899    0,019    0,898      0,899    0,897      0,879    0,986     0,942
```

FIGURA 56 - Salida Hoeffding Tree Fichero Maestro

Naïve Bayes

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      179734               66.5378 %
Incorrectly Classified Instances     90389               33.4622 %
Kappa statistic                          0.6127
Mean absolute error                      0.0654
Root mean squared error                  0.216
Relative absolute error                 42.5868 %
Root relative squared error             77.9318 %
Total Number of Instances           270123
```

```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j     k   <-- classified as
  1275   130   685     0  2094     0    16     0     0     0     0 |     a = 00000
  2012 15580  2470  3789  3707  3746  8552   152  7477     0     0 |     b = 00001
   358    52  5283     0     1     0   206     0     0     0     0 |     c = 00100
  1663  2158     0  9502  4562  1696  1371    14   277   357     0 |     d = 00101
  5309  3483    85  1167 40691     0  1038   131  5867    29     0 |     e = 00110
     0    84     0    84     0 51876    56     0     0     0     0 |     f = 00111
   396  5224  6956  1487  2683   538 19056    25  5560  1559     0 |     g = 01000
     0    84     0     7     2     0    16  3503     2     0     0 |     h = 01010
     0   295     0    65   311     0    37    10  2721     0     0 |     i = 01011
     8     3     0     5   225     0     0     0     0  7259     0 |     j = 10001
     0    10     0     3     0     0     0     0     0     0 22988 |     k = 11111
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
                 0,304    0,037    0,116      0,304   0,168      0,167    0,913     0,105     00000
                 0,328    0,052    0,575      0,328   0,418      0,350    0,826     0,525     00001
                 0,895    0,039    0,341      0,895   0,494      0,539    0,986     0,638     00100
                 0,440    0,027    0,590      0,440   0,504      0,473    0,902     0,524     00101
                 0,704    0,064    0,750      0,704   0,726      0,655    0,957     0,859     00110
                 0,996    0,027    0,897      0,996   0,944      0,931    0,998     0,991     00111
                 0,438    0,050    0,628      0,438   0,516      0,452    0,856     0,551     01000
                 0,969    0,001    0,913      0,969   0,941      0,940    1,000     0,957     01010
                 0,791    0,072    0,124      0,791   0,215      0,295    0,953     0,133     01011
                 0,968    0,007    0,789      0,968   0,869      0,870    0,998     0,942     10001
                 0,999    0,000    1,000      0,999   1,000      1,000    1,000     1,000     11111
Weighted Avg.    0,665    0,041    0,713      0,665   0,672      0,632    0,926     0,739
```

FIGURA 57 - Salida Naïve Bayes Fichero Maestro

Redes Bayesianas

```
=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      218704               80.9646 %
Incorrectly Classified Instances     51419               19.0354 %
Kappa statistic                          0.7767
Mean absolute error                      0.0399
Root mean squared error                  0.1595
Relative absolute error                 25.9924 %
Root relative squared error             57.566  %
Total Number of Instances           270123
```

```
=== Confusion Matrix ===

     a     b     c     d     e     f     g     h     i     j     k   <-- classified as
  3146    50    37     0   965     0     2     0     0     0     0 |     a = 00000
  1522 27596  1376  1929  3394  3347  6259     8  2053     1     0 |     b = 00001
    84   156  5434     7    10     0   209     0     0     0     0 |     c = 00100
    18  2116     0 17404   296   109  1391    11   181    74     0 |     d = 00101
   574  4047   416   174 47224     0   297    21  5042     5     0 |     e = 00110
     0    89     0    59     0 51816   136     0     0     0     0 |     f = 00111
    34  8000   914  1521   711     6 29299     7  2015   977     0 |     g = 01000
     0    40     0     0     3     0     2  3567     2     0     0 |     h = 01010
     0   305     0    15   171     0    27     0  2921     0     0 |     i = 01011
    13     9     0     4   169     0     9     0     0  7296     0 |     j = 10001
     0     0     0     0     0     0     0     0     0     0 23001 |     k = 11111
```

```
=== Detailed Accuracy By Class ===

               TP Rate   FP Rate   Precision   Recall  F-Measure   MCC      ROC Area  PRC Area  Class
               0,749     0,008     0,584      0,749     0,656      0,655     0,993     0,717     00000
               0,581     0,067     0,651      0,581     0,614      0,538     0,901     0,705     00001
               0,921     0,010     0,665      0,921     0,772      0,777     0,997     0,887     00100
               0,806     0,015     0,824      0,806     0,815      0,799     0,989     0,904     00101
               0,817     0,027     0,892      0,817     0,853      0,816     0,983     0,943     00110
               0,995     0,016     0,937      0,995     0,965      0,957     0,998     0,993     00111
               0,674     0,037     0,779      0,674     0,722      0,676     0,949     0,816     01000
               0,987     0,000     0,987      0,987     0,987      0,987     1,000     0,999     01010
               0,849     0,035     0,239      0,849     0,373      0,439     0,985     0,563     01011
               0,973     0,004     0,873      0,973     0,920      0,919     0,999     0,968     10001
               1,000     0,000     1,000      1,000     1,000      1,000     1,000     1,000     11111
Weighted Avg.  0,810     0,029     0,827      0,810     0,814      0,783     0,969     0,884
```

FIGURA 58 - Salida Redes Bayesianas Fichero Maestro

Conjunto 1

C4.5

```
Time taken to build model: 21.25 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      151857.0585          97.1034 %
Incorrectly Classified Instances      4529.9622           2.8966 %
Kappa statistic                          0.9681
Mean absolute error                      0.0039
Root mean squared error                  0.053
Relative absolute error                  4.0631 %
Root relative squared error             24.2197 %
Total Number of Instances           156387.0207
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,985    0,001    0,995      0,985    0,990      0,989    0,992     0,981     00000
                 0,938    0,007    0,930      0,938    0,934      0,928    0,993     0,931     00001
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00010
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00011
                 0,985    0,001    0,986      0,985    0,985      0,984    0,992     0,971     00100
                 0,981    0,003    0,969      0,981    0,975      0,972    0,996     0,979     00101
                 0,970    0,009    0,915      0,970    0,942      0,936    0,994     0,911     00110
                 0,988    0,002    0,982      0,988    0,985      0,984    1,000     0,998     00111
                 0,952    0,005    0,948      0,952    0,950      0,945    0,995     0,951     01000
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01001
                 0,951    0,001    0,991      0,951    0,971      0,968    0,976     0,952     01010
                 0,940    0,002    0,980      0,940    0,959      0,956    0,969     0,940     01011
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01100
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01101
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01110
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01111
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         10000
                 0,993    0,001    0,992      0,993    0,992      0,991    0,997     0,989     10001
                 1,000    0,000    0,999      1,000    1,000      1,000    1,000     1,000     11111
Weighted Avg.    0,971    0,003    0,972      0,971    0,971      0,968    0,991     0,964
```

```
=== Confusion Matrix ===

     a        b     c     d        e        f        g       h        i     j        k        l     m     n     o
 13996.98    27.08  0     0       94.78    10.15    84.63    0        0     0        0        0     0     0     0
     6.59 13337.67  0     0       61.08    54.49   115.27  235.03   348.2  0       41.32    16.77  0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
    50.6     79.52  0     0    13997.72     0        0       0       86.75  0        2.41     0     0     0     0
     0       71.08  0     0        0    13945.81    42.78   12.51   113.87  0       19.75     8.56  0     0     0
    15       82.89  0     0        0.25    32.71 13792.47   0       42.55  0       23.86   215.22  0     0     0
     0      168.37  0     0        0        5.73    0    14040.72    1.91  0        0        0.27  0     0     0
     0.33   251.75  0     0       38.25   237.36   53.29    2.62 13538.26  0       12.1     9.15  0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0      263.57  0     0        3.93    74.74   216.36   0       78.68  0    13516.77    23.6  0     0     0
     0       53.74  0     0        0       28.94   752.4    0        8.27  0       16.54 13357.12  0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
     0        1.9   0     0        1.9      3.79    24.64    0       66.35  0        1.9      0     0     0     0
     0        0     0     0        0        0        0       0        0     0        0        0     0     0     0
```

FIGURA 59 - Salida C4.5 Conjunto 1

Hoeffding Tree

```
Time taken to build model: 31.56 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      135983.0292          86.9529 %
Incorrectly Classified Instances     20403.9915          13.0471 %
Kappa statistic                          0.8565
Mean absolute error                      0.0166
Root mean squared error                  0.099
Relative absolute error                 17.3946 %
Root relative squared error             45.2571 %
Total Number of Instances           156387.0207

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,959    0,005    0,948      0,959    0,954      0,949    0,997     0,982     00000
               0,555    0,023    0,703      0,555    0,620      0,592    0,945     0,688     00001
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00010
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00011
               0,911    0,007    0,926      0,911    0,919      0,911    0,996     0,972     00100
               0,880    0,008    0,915      0,880    0,898      0,888    0,996     0,968     00101
               0,854    0,027    0,762      0,854    0,806      0,786    0,987     0,910     00110
               0,984    0,005    0,949      0,984    0,966      0,963    0,999     0,991     00111
               0,731    0,021    0,777      0,731    0,753      0,730    0,972     0,823     01000
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01001
               0,971    0,001    0,985      0,971    0,978      0,976    0,993     0,987     01010
               0,732    0,043    0,631      0,732    0,678      0,645    0,969     0,693     01011
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01100
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01101
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01110
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01111
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         10000
               0,988    0,002    0,980      0,988    0,984      0,982    0,999     0,994     10001
               1,000    0,000    0,998      1,000    0,999      0,999    1,000     1,000     11111
Weighted Avg.  0,870    0,013    0,870      0,870    0,868      0,856    0,987     0,910

=== Confusion Matrix ===

     a       b       c       d       e       f       g       h       i       j       k       l       m       n       o
 13634.78    44       0       0     179.4     0     328.34    0       0       0       0     27.08     0       0       0
   135.63  7888.29    0       0     393.41  611.67  394.01  750.3   1578.14   0      44.61  2420.35    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
   368.68  546.99    0       0    12956.75   0       2.41    0     342.17    0       0       0       0       0       0
     0     772.06    0       0       0    12515.56  32.91   5.27  564.07    0       4.61  297.5     0       0       0
   226.29  319.76    0       0       8.85    25.33 12139.55  0      22.14    0       6.15 1434.25    0       0       0
     0     183.65    0       0       0      47.21    0    13984.78  1.36    0       0       0       0       0       0
     0.65  769.64    0       0     453.8   429.94  116.72    0.33 10385.83   0      55.25 1781.54    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
    11.8   125.88    0       0       0      15.74   27.54    0      66.88    0   13811.81  129.82    0       0       0
     0     574.63    0       0       0      16.54 2819.42    0     310.05    0      90.95 10405.41    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       3.79    0       0       0       9.48   62.55    0      94.78    0       1.9     0       0       0       0
     0       0       0       0       0       0       0.62    0       0       0       0       0       0       0       0
```

FIGURA 60 - Salida Hoeffding Tree Conjunto 1

Naïve Bayes

```
Time taken to build model: 0.62 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      113595.9564          72.6377 %
Incorrectly Classified Instances     42791.0642          27.3623 %
Kappa statistic                          0.699
Mean absolute error                      0.0327
Root mean squared error                  0.1477
Relative absolute error                 34.1818 %
Root relative squared error             67.5131 %
Total Number of Instances           156387.0207

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,387    0,062    0,386      0,387    0,386      0,325    0,926     0,399     00000
               0,263    0,020    0,565      0,263    0,359      0,346    0,827     0,421     00001
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00010
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00011
               0,905    0,039    0,701      0,905    0,790      0,774    0,987     0,885     00100
               0,516    0,021    0,714      0,516    0,599      0,575    0,932     0,660     00101
               0,644    0,066    0,494      0,644    0,559      0,514    0,934     0,618     00110
               0,993    0,010    0,907      0,993    0,948      0,943    0,999     0,987     00111
               0,423    0,022    0,653      0,423    0,513      0,489    0,857     0,488     01000
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01001
               0,975    0,002    0,985      0,975    0,980      0,978    1,000     0,997     01010
               0,900    0,053    0,629      0,900    0,740      0,724    0,970     0,615     01011
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01100
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01101
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01110
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01111
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         10000
               0,985    0,007    0,938      0,985    0,961      0,957    0,999     0,990     10001
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     11111
Weighted Avg.  0,726    0,027    0,725      0,726    0,712      0,693    0,948     0,733

=== Confusion Matrix ===

    a       b       c   d   e       f       g       h       i       j   k       l       m   n   o
 5500.63  145.56   0   0   1655.26  50.77  6851.24  0      13.54   0   0       0       0   0   0
 1380.53 3735.61   0   0   1150.89 1704.78  937.42 1015.27 1530.23  0   63.17  2698.49  0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
 1101.22   96.39   0   0   12870    31.33   0       0     118.07   0   0       0       0   0   0
 3556.88  548.93   0   0      0    7337.55  280.39  368.59 1259.78  0   38.83   412.69  0   0   0
 2047.2   321.97   0   0     11.56  116.34 9151.28  0     129.38   0   18.45  2409.51  0   0   0
    0      33.56   0   0      0       35.2    0   14111.66  36.57   0   0       0       0   0   0
  580.66 1165.9    0   0   2669.2   758.19  416.86  67.35 6009.63   0   22.89  2020.54  0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
    3.93  180.96   0   0      0       62.94   11.8    0      78.68   0 13855.08   23.6    0   0   0
    0     388.6    0   0      0      173.63  764.8    0      20.67   0   70.28 12799.02  0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
   83.41    0      0   0      0       9.48  115.63    0       0    0   0       0       0   0   0
    0       0      0   0      0       0       0       0       0    0   0       0       0   0   0
```

FIGURA 61 - Salida Naïve Bayes Conjunto

Red Bayesiana

```
Bayes Network Classifier
not using ADTree
#attributes=16 #classindex=15
Network structure (nodes followed by parents)
ABL_E(18): WAVE
ABL_N(20): WAVE
ABL_Z(19): WAVE
COC_E(35): WAVE
COC_N(37): WAVE
COC_Z(35): WAVE
CON_E(27): WAVE
CON_N(62): WAVE
CON_Z(26): WAVE
LAR_E(80): WAVE
LAR_N(70): WAVE
LAR_Z(28): WAVE
SHA_E(59): WAVE
SHA_N(29): WAVE
SHA_Z(65): WAVE
WAVE(19):
LogScore Bayes: -6808225.218156283
LogScore BDeu: -6891399.89356155
LogScore MDL: -6886727.527727719
LogScore ENTROPY: -6815921.226754405
LogScore AIC: -6827244.226754405

Time taken to build model: 3.44 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      136053.5198          86.998  %
Incorrectly Classified Instances     20333.5009          13.002  %
Kappa statistic                          0.857
Mean absolute error                      0.0163
Root mean squared error                  0.102
Relative absolute error                 17.0166 %
Root relative squared error             46.6129 %
Total Number of Instances           156387.0207

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC Area  Class
               0,917    0,011    0,896      0,917   0,907      0,897    0,996     0,953     00000
               0,513    0,029    0,637      0,513   0,568      0,534    0,925     0,646     00001
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         00010
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         00011
               0,964    0,012    0,890      0,964   0,925      0,918    0,998     0,978     00100
               0,854    0,013    0,866      0,854   0,860      0,847    0,993     0,940     00101
               0,748    0,017    0,816      0,748   0,781      0,760    0,975     0,865     00110
               0,993    0,007    0,931      0,993   0,961      0,958    0,999     0,992     00111
               0,634    0,020    0,764      0,634   0,693      0,669    0,960     0,779     01000
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         01001
               0,995    0,001    0,994      0,995   0,995      0,994    1,000     1,000     01010
               0,962    0,031    0,759      0,962   0,848      0,839    0,993     0,926     01011
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         01100
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         01101
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         01110
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         01111
               0,000    0,000    0,000      0,000   0,000      0,000    ?         ?         10000
               0,989    0,003    0,969      0,989   0,979      0,977    1,000     0,995     10001
               1,000    0,000    1,000      1,000   1,000      1,000    1,000     1,000     11111
Weighted Avg.  0,870    0,013    0,866      0,870   0,865      0,854    0,985     0,916
```

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
13039.02	20.31	0	0	121.86	0	1035.81	0	0	0	0	0	0	0	0
651.19	7294.88	1.8	0	642.81	879.94	899.4	965.27	1799.99	0	14.67	1066.16	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
236.15	45.78	0	0	13698.92	16.87	12.05	0	207.23	0	0	0	0	0	0
46.07	839.86	0	0	0	12147.63	90.83	72.4	645.03	0	28.96	281.05	0	0	0
448.16	717	0	0	237.36	45.01	10632.99	0	74.77	0	19.43	2036.62	0	0	0
0	19.92	0	0	0	33.56	0	14123.94	39.57	0	0	0	0	0	0
64.74	2118.62	0	0	690.51	823.58	156.61	4.58	9011.99	0	18.31	954.03	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	39.34	0	0	0	11.8	3.93	0	3.93	0	14150.12	7.87	0	0	0
0	355.53	0	0	0	57.88	115.75	0	8.27	0	4.13	13675.44	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
60.66	3.79	0	0	0	5.69	81.51	0	3.79	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURA 62 - Salida Redes Bayesianas Conjunto 1

Conjunto 2

C4.5

```
Number of Leaves  :     1355

Size of the tree :      2709


Time taken to build model: 13.18 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      151240.7751          96.7093 %
Incorrectly Classified Instances      5146.2456           3.2907 %
Kappa statistic                          0.9638
Mean absolute error                      0.0045
Root mean squared error                  0.0562
Relative absolute error                  4.7166 %
Root relative squared error             25.6763 %
Total Number of Instances           156387.0207

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0,985    0,001    0,993      0,985    0,989      0,988    0,992     0,980     00000
               0,924    0,008    0,916      0,924    0,920      0,912    0,991     0,923     00001
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00010
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00011
               0,979    0,001    0,987      0,979    0,983      0,981    0,990     0,973     00100
               0,981    0,004    0,962      0,981    0,971      0,969    0,996     0,975     00101
               0,965    0,009    0,913      0,965    0,938      0,932    0,993     0,901     00110
               0,984    0,002    0,984      0,984    0,984      0,982    0,999     0,996     00111
               0,938    0,007    0,931      0,938    0,934      0,928    0,993     0,944     01000
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01001
               0,955    0,001    0,990      0,955    0,972      0,970    0,978     0,951     01010
               0,937    0,002    0,977      0,937    0,956      0,952    0,968     0,935     01011
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01100
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01101
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01110
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01111
               0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         10000
               0,991    0,001    0,992      0,991    0,992      0,991    0,996     0,987     10001
               1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     11111
Weighted Avg.  0,967    0,003    0,968      0,967    0,967      0,964    0,991     0,960

=== Confusion Matrix ===

     a       b       c       d       e       f       g       h       i       j       k       l       m       n       o
 14000.36   54.16   0       0       44      6.77    108.32  0       0       0       0       0       0       0       0
    11.68 13130.49  0       0       61.68   86.23   120.66  218.86  527.84  0       41.32   17.37   0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
    62.65   127.71  0       0    13918.2    0       0       0     103.62    0       4.82    0       0       0       0
     0       69.11  0       0       0    13953.71   46.07   13.16   113.21  0       13.82   6.58    0       0       0
    25.09   105.52  0       0       0.74    42.31 13714.99  0       45.01   0       26.56   248.43  0       0       0
     0      225.67  0       0       0       4.91    0    13984.78   1.09    0       0       0.55    0       0       0
     0      342.64  0       0       71.93   308.31   66.04   1.31 13329.67   0       11.44   7.52    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     3.93   212.43  0       0       0       78.68   188.83  0       94.41   0    13571.84   39.34   0       0       0
     0       70.28  0       0       0       24.8    744.13  0       28.94   0       24.8  13324.05  0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       3.79    3.79    30.33   0       72.03   0       11.37   0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
```

FIGURA 63 - Salida C4.5 Conjunto 2

Hoeffding Tree

```
Time taken to build model: 30.63 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      132288.9554          84.5908 %
Incorrectly Classified Instances     24098.0653          15.4092 %
Kappa statistic                          0.8305
Mean absolute error                      0.0205
Root mean squared error                  0.1065
Relative absolute error                 21.4686 %
Root relative squared error             48.6851 %
Total Number of Instances           156387.0207
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,939    0,009    0,910      0,939    0,924      0,917    0,994     0,964     00000
                 0,466    0,021    0,686      0,466    0,555      0,531    0,924     0,630     00001
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00010
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00011
                 0,897    0,014    0,868      0,897    0,883      0,871    0,994     0,942     00100
                 0,855    0,009    0,909      0,855    0,882      0,871    0,993     0,954     00101
                 0,754    0,023    0,769      0,754    0,761      0,738    0,979     0,868     00110
                 0,984    0,006    0,939      0,984    0,961      0,957    0,999     0,990     00111
                 0,668    0,018    0,787      0,668    0,723      0,700    0,960     0,789     01000
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01001
                 0,937    0,002    0,982      0,937    0,959      0,955    0,994     0,978     01010
                 0,816    0,065    0,555      0,816    0,661      0,634    0,963     0,629     01011
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01100
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01101
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01110
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01111
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         10000
                 0,988    0,002    0,979      0,988    0,984      0,982    0,999     0,994     10001
                 1,000    0,000    0,996      1,000    0,998      0,998    1,000     0,999     11111
Weighted Avg.    0,846    0,015    0,853      0,846    0,845      0,832    0,982     0,885
```

```
=== Confusion Matrix ===

     a       b    c    d       e       f       g       h       i    j       k       l    m    n    o
 13353.83   98.17  0    0     186.17    3.38  402.81    0       0    0       0     165.86  0    0    0
   161.98 6622.13  0    0     737.42  721.25  488.02  895.8  1336.52  0     104.79 3148.49  0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
   749.4   332.53 0    0   12759.15    0       0       0     375.91 0       0       0    0    0    0
     0     758.24 0    0       0   12161.45   34.88   14.48  596.98 0       2.63  643.71 0    0    0
   406.83  405.6  0    0      15.5    12.3  10717.36  0      11.31  0       3.69 2621.78 0    0    0
     0     175.73 0    0       0     48.03     0    13987.5   5.73  0       0       0    0    0    0
     0.98  784.02 0    0     992.29  414.57   81.41    1.31 9494.9  0      55.91 2139.22 0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0     141.62 0    0       3.93    7.87   39.34    0      62.94 0   13324.01  562.54 0    0    0
     0     330.72 0    0       0       0    2145.57    0      66.14 0     74.41 11600.15 0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0       0    0    0       0       0       0       0       0    0       0       0    0    0    0
     0       1.9  0    0       1.9     3.79   26.54    0     113.74 0       3.79   11.37 0    0    0
     0.62    0    0    0       0       0       0       0       0    0       0       0    0    0    0
```

FIGURA 64 - Salida Hoeffding Tree Conjunto 2

Naïve Bayes

```
Time taken to build model: 0.41 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      109684.9895          70.1369 %
Incorrectly Classified Instances     46702.0312          29.8631 %
Kappa statistic                          0.6715
Mean absolute error                      0.035
Root mean squared error                  0.1551
Relative absolute error                 36.6018 %
Root relative squared error             70.9116 %
Total Number of Instances           156387.0207
```

=== Detailed Accuracy By Class ===

	TP Rate	FP Rate	Precision	Recall	F-Measure	MCC	ROC Area	PRC Area	Class
	0,401	0,076	0,344	0,401	0,370	0,303	0,919	0,363	00000
	0,163	0,010	0,615	0,163	0,258	0,287	0,808	0,390	00001
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	00010
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	00011
	0,916	0,051	0,641	0,916	0,754	0,739	0,983	0,844	00100
	0,516	0,026	0,664	0,516	0,580	0,549	0,910	0,597	00101
	0,518	0,059	0,467	0,518	0,491	0,438	0,924	0,586	00110
	0,993	0,017	0,855	0,993	0,919	0,913	0,999	0,986	00111
	0,358	0,018	0,661	0,358	0,465	0,452	0,831	0,452	01000
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	01001
	0,931	0,001	0,985	0,931	0,957	0,954	0,999	0,994	01010
	0,934	0,064	0,594	0,934	0,726	0,715	0,962	0,528	01011
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	01100
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	01101
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	01110
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	01111
	0,000	0,000	0,000	0,000	0,000	0,000	?	?	10000
	0,985	0,005	0,954	0,985	0,969	0,966	0,999	0,991	10001
	1,000	0,000	1,000	1,000	1,000	1,000	1,000	1,000	11111
Weighted Avg.	0,701	0,030	0,707	0,701	0,681	0,665	0,940	0,703	

=== Confusion Matrix ===

a	b	c	d	e	f	g	h	i	j	k	l	m	n	o
5700.34	199.71	0	0	2003.92	0	6313.03	0	0	0	0	0	0	0	0
1871.25	2319.75	0	0	1589.22	1734.42	826.94	1155.68	1774.24	0	18.86	2926.34	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
1142.18	26.51	0	0	13021.81	0	0	0	26.51	0	0	0	0	0	0
3495.01	152.7	0	0	0	7330.96	652.93	1135.38	623.97	0	0	675.31	0	0	0
3421.92	27.3	0	0	5.17	119.54	7366.53	0	9.1	0	6.4	3236.95	0	0	0
0	23.47	0	0	0	55.39	0	14119.58	18.56	0	0	0	0	0	0
858.89	440.73	0	0	3692.55	1196.63	205.65	104.95	5092.22	0	58.85	2063.04	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
43.27	239.97	0	0	0	342.25	35.4	0	153.42	0	13233.53	165.22	0	0	0
0	334.86	0	0	0	268.71	219.1	0	0	0	115.75	13278.57	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
41.7	7.58	0	0	0	0	163.02	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

FIGURA 65 - Salida Naïve Bayes Conjunto 2

Red Bayesiana

```
Bayes Network Classifier
not using ADTree
#attributes=8 #classindex=7
Network structure (nodes followed by parents)
COC_Z(35): WAVE
CON_N(62): WAVE
CON_Z(26): WAVE
LAR_E(80): WAVE
LAR_N(70): WAVE
SHA_E(59): WAVE
SHA_Z(65): WAVE
WAVE(19):
LogScore Bayes: -3590630.9500695644
LogScore BDeu: -3648016.1248695026
LogScore MDL: -3644159.517619364
LogScore ENTROPY: -3597709.8838094384
LogScore AIC: -3605137.8838094384


Time taken to build model: 1.37 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      134655.6697          86.1041 %
Incorrectly Classified Instances     21731.351           13.8959 %
Kappa statistic                          0.8471
Mean absolute error                      0.0182
Root mean squared error                  0.1051
Relative absolute error                 18.994  %
Root relative squared error             48.0271 %
Total Number of Instances           156387.0207
```

```
=== Detailed Accuracy By Class ===

                 TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
                 0,910    0,012    0,886      0,910    0,898      0,888    0,996     0,948     00000
                 0,468    0,029    0,614      0,468    0,531      0,497    0,913     0,605     00001
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00010
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         00011
                 0,968    0,014    0,876      0,968    0,920      0,913    0,997     0,958     00100
                 0,854    0,016    0,842      0,854    0,848      0,833    0,992     0,931     00101
                 0,745    0,016    0,819      0,745    0,781      0,761    0,974     0,858     00110
                 0,991    0,008    0,925      0,991    0,957      0,953    0,999     0,989     00111
                 0,612    0,022    0,740      0,612    0,670      0,643    0,956     0,744     01000
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01001
                 0,970    0,003    0,975      0,970    0,973      0,970    1,000     0,996     01010
                 0,970    0,031    0,757      0,970    0,850      0,841    0,991     0,899     01011
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01100
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01101
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01110
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         01111
                 0,000    0,000    0,000      0,000    0,000      0,000    ?         ?         10000
                 0,983    0,002    0,977      0,983    0,980      0,978    1,000     0,996     10001
                 1,000    0,000    1,000      1,000    1,000      1,000    1,000     1,000     11111
Weighted Avg.    0,861    0,014    0,855      0,861    0,855      0,843    0,983     0,902
```

```
=== Confusion Matrix ===

     a       b       c       d       e       f       g       h       i       j       k       l       m       n       o
 12944.24   71.08    0       0      81.24    3.38  1117.05    0       0       0       0       0       0       0       0
   655.99 6650.57    0       0     743.41 1181.73   733.23 1088.02 2138.61    0     110.78  914.37    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
   216.87   65.06    0       0   13761.57    9.64    7.23    0     156.63    0       0       0       0       0       0
    20.4    803      0       0       0   12137.1    44.76   61.87   640.42    0      33.57  460.74    0       0       0
   570.16  479.39    0       0     101.83   68.63 10597.57    0      25.58    0      49.93 2310.14    0       0       0
     0      22.65    0       0       0       60.03    0    14089.56  44.75    0       0       0       0       0       0
    94.49 2308.25    0       0    1023.67  853.66  171.32    0.65  8694.86    0     128.16  649.97    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     7.87  220.3     0       0       0      15.74   39.34    0      35.4     0   13796.08   90.48    0       0       0
     0     202.57    0       0       0      78.55  111.62    0       4.13    0      33.07 13787.06    0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
   100.47    1.9     0       0       0       5.69  115.63    0      13.27    0       0       0       0       0       0
     0       0       0       0       0       0       0       0       0       0       0       0       0       0       0
```

FIGURA 66 - Salida Redes Bayesianas Conjunto 2

Segmentación

Canopy con 3 clústeres, 1 semilla

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2  <-- assigned to cluster
  2721     0  1479 | 00000
 32175  3935 11375 | 00001
  4585     0  1315 | 00100
 17715  2011  1874 | 00101
 30196   246 27358 | 00110
  7587 44513     0 | 00111
 35530   185  7769 | 01000
  3494    60    60 | 01010
  3200   239     0 | 01011
  7315     0   185 | 10001
 20147     0  2854 | 11111

Cluster 0 <-- 01000
Cluster 1 <-- 00111
Cluster 2 <-- 00110

Incorrectly clustered instances :       162722.0        60.24   %

       === Model and evaluation on training set ===

       Clustered Instances

       0      164665 ( 61%)
       1       51189 ( 19%)
       2       54269 ( 20%)
```

FIGURA 67 - Salida Canopy 3 Clústeres, 1 Semilla

Canopy con 5 clústeres, 1 semilla

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4  <-- assigned to cluster
  2721     0  1479     0     0 | 00000
 31037  3450 11375     0  1623 | 00001
  4585     0  1315     0     0 | 00100
 17713  1840  1874     0   173 | 00101
 30115   233 27358     0    94 | 00110
  7587 44336     0     0   177 | 00111
 35286   120  7769     0   309 | 01000
  2355    38    39  1133    49 | 01010
  3086   217     0     0   136 | 01011
  7315     0   185     0     0 | 10001
 20147     0  2854     0     0 | 11111

Cluster 0 <-- 01000
Cluster 1 <-- 00111
Cluster 2 <-- 00110
Cluster 3 <-- 01010
Cluster 4 <-- 00001

Incorrectly clustered instances :       160387.0        59.3755 %
```

```
=== Model and evaluation on training set ===

Clustered Instances

0      161947 ( 60%)
1       50234 ( 19%)
2       54248 ( 20%)
3        1133 (  0%)
4        2561 (  1%)
```

FIGURA 68 – Salida Canopy 5 Clústeres, 1 Semilla

Canopy con 3 clústeres, 3 semillas

```
Class attribute: WAVE
Classes to Clusters:

    0     1     2  <-- assigned to cluster
  313  3180   707 | 00000
 3142 29204 15139 | 00001
  742  5158     0 | 00100
    0 11452 10148 | 00101
  498 37727 19575 | 00110
    0  4271 47829 | 00111
 1898 32224  9362 | 01000
  839  1580  1195 | 01010
    0  2353  1086 | 01011
    0  7500     0 | 10001
    0 23001     0 | 11111

Cluster 0 <-- 00001
Cluster 1 <-- 00110
Cluster 2 <-- 00111

Incorrectly clustered instances :       181425.0        67.1638 %

        === Model and evaluation on training set ===

        Clustered Instances

        0        7432 (  3%)
        1      157650 ( 58%)
        2      105041 ( 39%)
```

FIGURA 69 - Salida Canopy 3 Clústeres, 3 Semillas

Canopy con 5 clústeres, 5 semillas

```
Class attribute: WAVE
Classes to Clusters:

    0     1     2     3     4  <-- assigned to cluster
  253   563     0  2677   707 | 00000
 2358  2357  2725 25612 14433 | 00001
  677  1717     0  3506     0 | 00100
    0    54   431 11356  9759 | 00101
  351   897   314 36808 19430 | 00110
    0     0  6576  2955 42569 | 00111
 1851   800   662 30955  9216 | 01000
  730   298   172  1375  1039 | 01010
    0     0   246  2265   928 | 01011
    0     0     4  7496     0 | 10001
    0     0     0 23001     0 | 11111

Cluster 0 <-- 01000
Cluster 1 <-- 00100
Cluster 2 <-- 00001
Cluster 3 <-- 00110
Cluster 4 <-- 00111

Incorrectly clustered instances :       184453.0        68.2848 %

          === Model and evaluation on training set ===

          Clustered Instances

          0       6220 (  2%)
          1       6686 (  2%)
          2      11130 (  4%)
          3     148006 ( 55%)
          4      98081 ( 36%)
```

FIGURA 70 - Salida Canopy 5 Clústeres, 5 Semillas

LVQ con razón de aprendizaje 1.0 y 11 Clústeres

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
     0  2131     0     0     0   317   551     0     0     1  1200 | 00000
     0 18955    94  2425     0  1459 12466     0   228  5422  6436 | 00001
     0  2292     0     0     0  1428     0     0    33     1  2146 | 00100
     0  8530     0   380     0     5  8445     0     0  3754   486 | 00101
     0 30032     0     8     0   688 18024     0     3  2587  6458 | 00110
     0   568     0  6321     0     0 38350     0     0  6704   157 | 00111
     0 24305    11   641     0   734  7208     0    10  4256  6319 | 01000
    41   636   245     7   160   146   519     5    82   433  1340 | 01010
     0  1783     0     9     0     0   839     0     0   504   304 | 01011
     0  6878     0     0     0     0     0     0     0     4   618 | 10001
     0 21243     0     0     0     0     0     0     0   145  1613 | 11111

Cluster  0 <-- No class
Cluster  1 <-- 00110
Cluster  2 <-- 01010
Cluster  3 <-- 00001
Cluster  4 <-- No class
Cluster  5 <-- 00100
Cluster  6 <-- 00111
Cluster  7 <-- No class
Cluster  8 <-- No class
Cluster  9 <-- 00101
Cluster 10 <-- 01000

Incorrectly clustered instances :       187570.0         69.4387 %


        Time taken to build model (full training data) : 376.67 seconds

        === Model and evaluation on training set ===

        Clustered Instances

         0          41 (  0%)
         1      117353 ( 43%)
         2         350 (  0%)
         3        9791 (  4%)
         4         160 (  0%)
         5        4777 (  2%)
         6       86402 ( 32%)
         7           5 (  0%)
         8         356 (  0%)
         9       23811 (  9%)
        10       27077 ( 10%)
```

FIGURA 71 - LVQ con razón de aprendizaje 1.0 y 11 Clústeres

LVQ con razón de aprendizaje 0.8 y 11 Clústeres

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
  2114     0     3     0     0     0     0     0   159  1921     3 | 00000
 16939     0  8400   487   217  2375  8152     0  1054  2380  7481 | 00001
   140     0  2641     0     0     0     0     0    23  3096     0 | 00100
  8585     0    88   736     0   249  7873     0    39   136  3894 | 00101
 34817     0  1953   202     0   261 13565     0   592  1761  4649 | 00110
   611     0     0  5441     0  4318 39401     0     0     0  2329 | 00111
 11665     0 12198    99   537   567  5981     0   120  1495 10822 | 01000
  2299   262     0    51   479    44    54    22   217     5   181 | 01010
  1260     0     0     6     0   135   879     0     0     0  1159 | 01011
  7157     0   101     0     0     0     0     0    58    90    94 | 10001
 20163     0   264     0     0     0     0     0    44    80  2450 | 11111

Cluster  0 <-- 00110
Cluster  1 <-- No class
Cluster  2 <-- 01000
Cluster  3 <-- 00101
Cluster  4 <-- 01010
Cluster  5 <-- 01011
Cluster  6 <-- 00111
Cluster  7 <-- No class
Cluster  8 <-- 00000
Cluster  9 <-- 00100
Cluster 10 <-- 00001

Incorrectly clustered instances :       171621.0        63.5344 %

       Time taken to build model (full training data) : 355.08 seconds

       === Model and evaluation on training set ===

       Clustered Instances

        0      105750 ( 39%)
        1         262 (  0%)
        2       25648 (  9%)
        3        7022 (  3%)
        4        1233 (  0%)
        5        7949 (  3%)
        6       75905 ( 28%)
        7          22 (  0%)
        8        2306 (  1%)
        9       10964 (  4%)
       10       33062 ( 12%)
```

FIGURA 72 - LVQ con razón de aprendizaje 0.8 y 11 Clústeres

SKMeans: distancia Euclídea

```
Class attribute: WAVE
Classes to Clusters:

    0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
    0  1940     0     0     0     0     0   708     0  1552     0 | 00000
 9592  7450   816  1392   295  8692  2789  3523  2247 10444   245 | 00001
    0     0     0     0     0     0     0   658     0  5242     0 | 00100
 2833   147    57  1714     0  2396  3977  6390  4023    63     0 | 00101
 7939 29199     0     0     0  5857     0 10273     0  4532     0 | 00110
    0     0 13765 16543 18628     2  1234     0  1928     0     0 | 00111
 5325  5622     0     0    22  3906  5517  3845  3871 14401   975 | 01000
  289    20     0     3     0   365     0    82     0    19  2836 | 01010
 1950     0     0     0    21  1468     0     0     0     0     0 | 01011
    0     0     0     0     0     4     0  6505     0   991     0 | 10001
    0    82     0     0     0     0     0 22744     0   175     0 | 11111

Cluster  0 <-- 00001
Cluster  1 <-- 00110
Cluster  2 <-- No class
Cluster  3 <-- No class
Cluster  4 <-- 00111
Cluster  5 <-- 01011
Cluster  6 <-- No class
Cluster  7 <-- 11111
Cluster  8 <-- 00101
Cluster  9 <-- 01000
Cluster 10 <-- 01010

Incorrectly clustered instances :       167232.0        61.9096 %

        Time taken to build model (full training data) : 39.39 seconds

        === Model and evaluation on training set ===

        Clustered Instances

         0       27928 ( 10%)
         1       44460 ( 16%)
         2       14638 (  5%)
         3       19652 (  7%)
         4       18966 (  7%)
         5       22690 (  8%)
         6       13517 (  5%)
         7       54728 ( 20%)
         8       12069 (  4%)
         9       37419 ( 14%)
        10        4056 (  2%)
```

FIGURA 73 - 500 iteraciones, 11 semillas

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
  1937     0   709  1554     0     0     0     0     0     0     0 | 00000
  7445   489  3514 10470  2241  5748  9340  4523  1971  1499   245 | 00001
     0     0   657  5243     0     0     0     0     0     0     0 | 00100
   148     0  6389    63  1264  2299  1726  1857  6769  1085     0 | 00101
 29181     0 10285  4534     0  3221  6637  3942     0     0     0 | 00110
     0 18242     0     0 10173   185     0     1  1016 22483     0 | 00111
  5531    32  3832 14410    97  2577  5596  2697  7728     0   984 | 01000
    22     0    73    18     1   279   196   201     0     2  2822 | 01010
     0    10     0     0     0   981  1464   984     0     0     0 | 01011
     0     0  6500   998     0     2     0     0     0     0     0 | 10001
    89     0 22735   177     0     0     0     0     0     0     0 | 11111

Cluster  0 <-- 00110
Cluster  1 <-- No class
Cluster  2 <-- 11111
Cluster  3 <-- 01000
Cluster  4 <-- No class
Cluster  5 <-- 10001
Cluster  6 <-- 00001
Cluster  7 <-- 01011
Cluster  8 <-- 00101
Cluster  9 <-- 00111
Cluster 10 <-- 01010

Incorrectly clustered instances :       161397.0        59.7494 %
```

```
=== Model and evaluation on training set ===

Clustered Instances

 0       44353 ( 16%)
 1       18773 (  7%)
 2       54694 ( 20%)
 3       37467 ( 14%)
 4       13776 (  5%)
 5       15292 (  6%)
 6       24959 (  9%)
 7       14205 (  5%)
 8       17484 (  6%)
 9       25069 (  9%)
10        4051 (  1%)
```

FIGURA 74 - 400 iteraciones, 11 semillas

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
  1937     0   709  1554     0     0     0     0     0     0     0 | 00000
  7445   489  3514 10470  2241  5748  9340  4523  1971  1499   245 | 00001
     0     0   657  5243     0     0     0     0     0     0     0 | 00100
   148     0  6389    63  1264  2299  1726  1857  6769  1085     0 | 00101
 29181     0 10285  4534     0  3221  6637  3942     0     0     0 | 00110
     0 18242     0     0 10173   185     0     1  1016 22483     0 | 00111
  5531    32  3832 14410    97  2577  5596  2697  7728     0   984 | 01000
    22     0    73    18     1   279   196   201     0     2  2822 | 01010
     0    10     0     0     0   981  1464   984     0     0     0 | 01011
     0     0  6500   998     0     2     0     0     0     0     0 | 10001
    89     0 22735   177     0     0     0     0     0     0     0 | 11111

Cluster  0 <-- 00110
Cluster  1 <-- No class
Cluster  2 <-- 11111
Cluster  3 <-- 01000
Cluster  4 <-- No class
Cluster  5 <-- 10001
Cluster  6 <-- 00001
Cluster  7 <-- 01011
Cluster  8 <-- 00101
Cluster  9 <-- 00111
Cluster 10 <-- 01010

Incorrectly clustered instances :       161397.0        59.7494 %

                === Model and evaluation on training set ===

                Clustered Instances

                 0       44353 ( 16%)
                 1       18773 (  7%)
                 2       54694 ( 20%)
                 3       37467 ( 14%)
                 4       13776 (  5%)
                 5       15292 (  6%)
                 6       24959 (  9%)
                 7       14205 (  5%)
                 8       17484 (  6%)
                 9       25069 (  9%)
                10        4051 (  1%)
```

FIGURA 75 - 300 iteraciones , 11 semillas

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
  1937     0   709  1554     0     0     0     0     0     0     0 | 00000
  7445   489  3514 10470  2241  5748  9340  4523  1971  1499   245 | 00001
     0     0   657  5243     0     0     0     0     0     0     0 | 00100
   148     0  6389    63  1264  2299  1726  1857  6769  1085     0 | 00101
 29181     0 10285  4534     0  3221  6637  3942     0     0     0 | 00110
     0 18242     0     0 10173   185     0     1  1016 22483     0 | 00111
  5531    32  3832 14410    97  2577  5596  2697  7728     0   984 | 01000
    22     0    73    18     1   279   196   201     0     2  2822 | 01010
     0    10     0     0     0   981  1464   984     0     0     0 | 01011
     0     0  6500   998     0     2     0     0     0     0     0 | 10001
    89     0 22735   177     0     0     0     0     0     0     0 | 11111

Cluster  0 <-- 00110
Cluster  1 <-- No class
Cluster  2 <-- 11111
Cluster  3 <-- 01000
Cluster  4 <-- No class
Cluster  5 <-- 10001
Cluster  6 <-- 00001
Cluster  7 <-- 01011
Cluster  8 <-- 00101
Cluster  9 <-- 00111
Cluster 10 <-- 01010

Incorrectly clustered instances :       161397.0        59.7494 %
```

```
=== Model and evaluation on training set ===

Clustered Instances

 0       44353 ( 16%)
 1       18773 (  7%)
 2       54694 ( 20%)
 3       37467 ( 14%)
 4       13776 (  5%)
 5       15292 (  6%)
 6       24959 (  9%)
 7       14205 (  5%)
 8       17484 (  6%)
 9       25069 (  9%)
10        4051 (  1%)
```

FIGURA 76 - 200 iteraciones, 11 semillas

SKMeans: distancia Manhattan

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
   810     0  1508  1882     0     0     0     0     0     0     0 | 00000
  3232  1409  3991 14360   503  3568  4527  6390  1647  1819  6039 | 00001
     0     0   336  5563     0     1     0     0     0     0     0 | 00100
     0    27  6565    35     3     0  5357  1365  7914   135   199 | 00101
 17811     0 19839  6348     0   235    92  6841     0     0  6634 | 00110
     0 20550     0     0 15132    14   384     0   376 15644     0 | 00111
  1681     0  4923 18267     0  5653   215  3243  6523     0  2979 | 01000
     0     0   137   773     0    12   332   493     0    18  1849 | 01010
     0     9     0     0     0    85    34  1763     0     0  1548 | 01011
     0     0  6450   969     0     0     0     0     0     0    81 | 10001
     0     0 22992     9     0     0     0     0     0     0     0 | 11111

Cluster  0 <-- 00110
Cluster  1 <-- 00111
Cluster  2 <-- 11111
Cluster  3 <-- 01000
Cluster  4 <-- No class
Cluster  5 <-- 01011
Cluster  6 <-- No class
Cluster  7 <-- 00001
Cluster  8 <-- 00101
Cluster  9 <-- No class
Cluster 10 <-- 01010

Incorrectly clustered instances :       174265.0        64.5132 %

                === Model and evaluation on training set ===

                Clustered Instances

                 0       23534 (  9%)
                 1       21995 (  8%)
                 2       66741 ( 25%)
                 3       48206 ( 18%)
                 4       15638 (  6%)
                 5        9568 (  4%)
                 6       10941 (  4%)
                 7       20095 (  7%)
                 8       16460 (  6%)
                 9       17616 (  7%)
                10       19329 (  7%)
```

FIGURA 77 - 500 iteraciones, 11 semillas

```
Class attribute: WAVE
Classes to Clusters:

     0     1     2     3     4     5     6     7     8     9    10  <-- assigned to cluster
   810     0  1508  1882     0     0     0     0     0     0     0 | 00000
  3232  1409  3991 14360   503  3568  4527  6390  1647  1819  6039 | 00001
     0     0   336  5563     0     1     0     0     0     0     0 | 00100
     0    27  6565    35     3     0  5357  1365  7914   135   199 | 00101
 17811     0 19839  6348     0   235    92  6841     0     0  6634 | 00110
     0 20550     0     0 15132    14   384     0   376 15644     0 | 00111
  1681     0  4923 18267     0  5653   215  3243  6523     0  2979 | 01000
     0     0   137   773     0    12   332   493     0    18  1849 | 01010
     0     9     0     0     0    85    34  1763     0     0  1548 | 01011
     0     0  6450   969     0     0     0     0     0     0    81 | 10001
     0     0 22992     9     0     0     0     0     0     0     0 | 11111

Cluster  0 <-- 00110
Cluster  1 <-- 00111
Cluster  2 <-- 11111
Cluster  3 <-- 01000
Cluster  4 <-- No class
Cluster  5 <-- 01011
Cluster  6 <-- No class
Cluster  7 <-- 00001
Cluster  8 <-- 00101
Cluster  9 <-- No class
Cluster 10 <-- 01010

Incorrectly clustered instances :       174265.0        64.5132 %

                === Model and evaluation on training set ===

                Clustered Instances

                 0       23534 (  9%)
                 1       21995 (  8%)
                 2       66741 ( 25%)
                 3       48206 ( 18%)
                 4       15638 (  6%)
                 5        9568 (  4%)
                 6       10941 (  4%)
                 7       20095 (  7%)
                 8       16460 (  6%)
                 9       17616 (  7%)
                10       19329 (  7%)
```

FIGURA 78 - 400 iteraciones, 11 semillas

ANEXO C: Planificación

Id.	Nombre de tarea	Comienzo	Fin	Duración
1	Fase Inicial	05/05/2017	14/05/2017	10d
2	Investigación	01/06/2017	17/06/2017	17d
3	Experimentación	03/07/2017	22/07/2017	20d
4	Elaboración de la Memoria	07/08/2017	13/08/2017	7d
5	Contratiempos	16/08/2017	24/08/2017	9d
6	Revisión General	01/09/2017	05/09/2017	5d
7	Entrega	26/09/2017	26/09/2017	1d

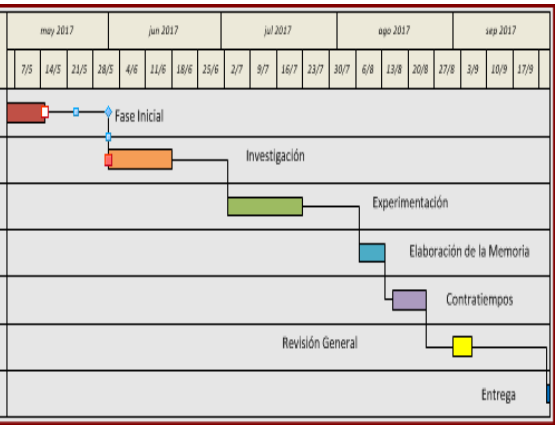


Tabla 24 - Planificación Inicial

Id.	Nombre de tarea	Comienzo	Fin	Duración
1	Fase Inicial	05/05/2017	14/05/2017	10d
2	Establecer Objetivos	05/05/2017	08/05/2017	4d
3	Adquisición de Datos	08/05/2017	11/05/2017	4d
4	Entendimiento de los Datos	11/05/2017	14/05/2017	4d
5	Investigación	24/06/2017	14/07/2017	21d
6	Contratiempos	24/06/2017	29/08/2017	67d
7	Programación	27/06/2017	18/07/2017	22d
8	Experimentación	18/07/2017	25/08/2017	39d
9	Clasificación	18/07/2017	06/08/2017	20d
10	Elaboración de la Memoria	25/06/2017	15/09/2017	83d
11	Segmentación	09/08/2017	24/08/2017	16d
12	Revisión General	15/09/2017	25/09/2017	11d
13	Entrega	26/09/2017	26/09/2017	1d

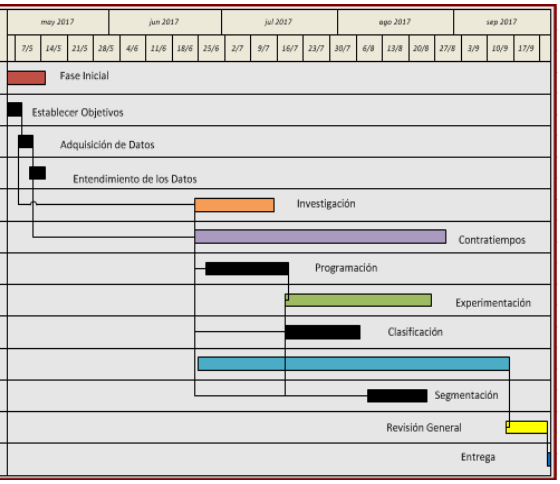


Tabla 25 - Planificación Final